Kernel Machine Methods for Risk Prediction with High Dimensional Data

Dissertation Advisor: Professor Tianxi Cai Jennifer Anne Sinnott

Kernel Machine Methods for Risk Prediction with High Dimensional Data

Abstract

Understanding the relationship between genomic markers and complex disease could have a profound impact on medicine, but the large number of potential markers can make it hard to differentiate true biological signal from noise and false positive associations.

A standard approach for relating genetic markers to complex disease is to test each marker for its association with disease outcome by comparing disease cases to healthy controls. It would be cost-effective to use control groups across studies of many different diseases; however, this can be problematic when the controls are genotyped on a platform different from the one used for cases. Since different platforms genotype different SNPs, imputation is needed to provide full genomic coverage, but introduces differential measurement error. In Chapter 1, we consider the effects of this differential error on association tests. We quantify the inflation in Type I Error by comparing two healthy control groups drawn from the same cohort study but genotyped on different platforms, and assess several methods for mitigating this error.

Analyzing genomic data one marker at a time can effectively identify associations, but the resulting lists of significant SNPs or differentially expressed genes can be hard to interpret. Integrating prior biological knowledge into risk prediction with such data by grouping genomic features into pathways reduces the dimensionality of the problem and could improve models by making them more biologically grounded and interpretable. The kernel machine framework has been proposed to model pathway effects because it allows nonlinear associations between the genes in a pathway and disease risk. In Chapter 2, we propose kernel machine regression under the accelerated failure time model. We derive a pseudo-score statistic for testing and a risk score for prediction using genes in a single pathway. We propose omnibus procedures that alleviate the need to prespecify the kernel and allow the data to drive the complexity of the resulting model. In Chapter 3, we extend methods for risk prediction using a single pathway to methods for risk prediction model using multiple pathways using a multiple kernel learning approach to select important pathways and efficiently combine information across pathways.

Contents

Acknowledgments

Thank you to my advisor, Tianxi Cai, for being the most amazing advisor I could have imagined. She is so supportive and has pushed me so far, and she constantly inspires me with her tirelessness and her infectious enthusiasm. I feel so lucky that we get to work together.

Thank you to my committee: to Peter Kraft, for always asking me interesting but difficult questions, and believing in my ability to answer them; and to Lorelei Mucci, for being such a wonderful collaborator and friend, and making me happier when I run into her in the stairwells.

Thank you to my dear friends: to Ravi Goyal, for making it through with me – I feel so proud of everything we've accomplished in the past four years, not to mention the past four months! Thank you to Stacey Ackerman-Alexeeff and Nathan Stein for countless hours in libraries and coffee shops. Thank you to Alexa Photopoulos, for being such a wonderful and supportive friend since that first day of college.

Finally, thank you to my family: to Sean Sinnott, for being an awesome big brother; and to my incredible parents, Warren and Loraine Sinnott, without whom I would be nothing, both literally and – more importantly – figuratively. I am so grateful for everything you have done for me.

Artifact due to differential error when cases and controls are imputed from different platforms

Jennifer A. Sinnott and Peter Kraft

Department of Biostatistics

Harvard School of Public Health

Abstract

Including previously-genotyped controls in a genome-wide association study can provide cost-savings, but can also create design biases. When cases and controls are genotyped on different platforms, the imputation needed to provide genome-wide coverage will introduce differential measurement error and may lead to false positives. We compared genotype frequencies of two healthy control groups from the Nurses' Health Study genotyped on different platforms (Affymetrix 6.0 [n=1,672] and Illumina HumanHap550 [n=1,038]). Using standard imputation quality filters, we observed 9,841 SNPs out of 2,347,809 (0.4%) significant at the 5×10^{-8} level. We explored three methods for controlling for this Type I error inflation. One method was to remove platform effects using principal components; another was to restrict to SNPs of highest quality imputation; and a third was to genotype some controls alongside cases to exclude SNPs that are statistical artifact. The first method could not reduce the Type I error rate; the other two could dramatically reduce the error rate, although both required that a portion of SNPs be excluded from analysis. Ideally, the biases we describe would be eliminated at the design stage, by genotyping sufficient numbers of cases and controls on each platform. Researchers using imputation to combine samples genotyped on different platforms with severely unbalanced case-control ratios should be aware of the potential for inflated Type I error rates and apply appropriate quality filters. Every SNP found with genome-wide significance should be validated on another platform to verify that its significance is not an artifact of study design.

1.1 Introduction

A population-based genome-wide association (GWA) study requires thousands of cases and controls in order to detect moderate associations between SNPs and disease, and each person genotyped can cost hundreds of dollars. Thus, when researchers plan numerous GWA studies for different diseases, it would be attractive to use the same healthy control group for more than one disease if all cases are being drawn from the same underlying population. The Wellcome Trust Case Control Consortium (WTCCC) demonstrated the effectiveness of this approach by comparing case groups of 7 major diseases to a shared control group (Wellcome Trust Case Control Consortium, 2007). Additionally, researchers may want to bring in publicly available controls to increase power without increasing cost. Zhuang et al. (2010) advocated this approach, and Ho and Lange (2010) did extensive simulations in this vein that demonstrate the potential improvement in power. Ho and Lange provided some examples of studies that have augmented their control groups with publicly available controls (Hom et al., 2008; Wrensch et al., 2009).

A complication in the reuse of control groups or the inclusion of external controls arises when investigators wish to genotype cases on a platform different from the one used for controls. This may easily happen as genotyping technology changes and new chips with new pricing plans become available. It can appear necessary when funding is too limited to support a sufficiently powered study with both cases and controls genotyped together. Moreover, even if funding exists to genotype or re-genotype a control group on a particular chip, there may be limited biological samples available for use, or a desire to conserve such samples. However, while each platform genotypes a collection of tagging SNPs, different platforms choose these tagging SNPs in different ways. For example, Illumina uses patterns of linkage disequilibrium in the HapMap to choose its tagging SNPs, while Affymetrix (Affy) provides a large but less determinate collection of SNPs designed to give good coverage of the entire genome. There is not necessarily much overlap between the SNPs genotyped on two different platforms. For example, there were 140,325 SNPs in the overlap between the 508,123 markers on the Illumina HumanHap550 chip and the 606,625 markers on the Affymetrix Genome-Wide Human 6.0 array we use in this study. Thus, if we restricted to SNPs in the overlap, we would drop about three-quarters of the SNPs we have available on each of these chips.

When pooling genotype data from different platforms, investigators could impute the SNPs missing on each platform to get a data set with comparable variables. This approach has been suggested as a way of combining study cases and controls with publicly available controls genotyped on a different chip (Zhuang

et al., 2010). Fallin et al. (2010) used imputation to combine their case-control study, genotyped on Illumina, with a publicly available case-control study genotyped on Affy. A number of imputation methods exist, and they have been shown to be very accurate in the typical setting where cases and controls are genotyped together on the same platform (Li et al., 2010; Howie et al., 2009). However, their performance in the setting we are discussing here, when cases and controls have been genotyped on different platforms, has been largely unexplored.

After imputation, investigators run association tests as usual, producing p-values for each SNP and looking for the most significant SNPs. However, the imputation has introduced differential measurement error: for example, some SNPs are measured almost perfectly (through actual genotyping) among the controls, but measured imperfectly (through imputation based on nearby measured SNPs) among the cases. Furthermore, the imputation itself may introduce bias. Many imputation programs base the imputation on a database of known genomes, such as the HapMap. If the minor allele frequency (MAF) of a SNP in the HapMap differs substantially from the MAF in study data, imputation in cases only or controls only can yield very different MAFs in cases and controls. This setting has been recognized as potentially problematic. For example, when discussing combining data from studies using different genotyping platforms, Li et al. (2010) recommends imputing and doing association tests within platform and then combining the results using a meta-analysis approach, which cannot be implemented unless each platform has at least some cases and controls.

Differential error induced by imputation may yield SNPs that appear to differ substantially between cases and controls purely as a result of the imputation. Past studies have shown that differential genotyping error between cases and controls can inflate Type I error rates (e.g. Moskvina et al., 2006). A recent study by Sebastiani et al. (2010) which built a model using 150 SNPs to predict longevity has been criticized for not controlling for different chips used with different frequencies between cases and controls. Critics suspect that many of the significant SNPs it identified are artifact of differential genotyping errors between these different chips (Alberts, 2010; Carmichael, 2010).

In this paper, we are concerned with problems occurring one step further down the pipeline. Under the assumption that markers actually genotyped by each chip are being genotyped with good accuracy, we investigate how well Type I error rates are maintained after imputation in a study where cases and controls are genotyped on different platforms. To do this, we used the healthy control groups from two studies nested within the Nurses' Health Study: a Type 2 Diabetes (T2D) study genotyped on Affy, and a

Breast Cancer (BrCa) study genotyped on Illumina. After imputation within each study, we label the T2D controls "cases" and the BrCa controls "controls," and fit a logistic regression predicting this case-control status from each SNP. We expect there to be no substantial genetic differences between these two groups – so any significant differences we see reflect a Type I error rate higher than expected.

When we did in fact observe inflated Type I error after applying standard imputation quality filters, we explored a number of ways to lessen the inflation. We first considered controlling for platform effect as we would control for population stratification: by using principal components (PCs) as covariates in logistic regression. However, the platform effect was so strong and confounded with case-control status that we could not fit the models. Then we considered restricting to SNPs imputed with good accuracy. This approach yields excellent results, but reduces power by reducing the number of SNPs we can test. Finally, we considered the possibility of genotyping a small number of additional controls alongside cases on the new platform, who could be compared to the original controls in a preliminary analysis to identify aberrant SNPs. This approach yields good results, but requires the additional expense of genotyping more subjects.

1.2 Methods

The BrCa and T2D studies have been described elsewhere (Hunter et al., 2007; Qi et al., 2010). Both studies were restricted to women of European ancestry. Genotyping in the BrCa study was done on the Illumina HumanHap550 chip, while the T2D study was genotyped on the Affymetrix Genome-Wide Human 6.0 array. We imputed missing genotypes separately within each study using MaCH 1.0, which relies on Markov chain haplotyping (`http://www.sph.umich.edu/csg/yli/mach/index.html`) (Li et al., 2009, 2010). We present results from imputation done separately in the two studies; when the two control groups were pooled first and then the imputation was done, results were similar. The imputations used HapMap Release 22 (NCBI build 36) as a reference panel. For each unmeasured SNP, we considered both a *soft call*, or *dosage*, imputation, which gives the expected number of rare alleles given the other SNPs available for that individual and takes values on a continuum between 0 and 2, and a *hard call* imputation, which gives the best integral guess for the number of rare alleles, either 0, 1, or 2. We had available 1,038 BrCa controls, which we labeled "controls," and 1,672 T2D controls, which we labeled "cases." SNPs with MAF < 0.025 (calculated using both groups after imputation) or imputation quality $R^2 < 0.30$ (calculated in either group) were removed.

We ran a logistic regression for each of m SNPs, modeling the log-odds of being a "case" ($Y = 1$) as a linear function of the number of rare alleles at the locus. That is, for the i^{th} SNP, $i = 1, \ldots, m$, with A_i copies of the rare allele, we fit

$$\log\left\{\frac{P(Y_i = 1)}{1 - P(Y_i = 1)}\right\} = \beta_0 + \beta_1 A_i$$

where β_1 is the effect of SNP i and β_0 is an intercept term. We stored the p-value and the χ^2 test statistic for the Wald test of β_1. For the soft call genotypes, where A_i is the expected number of rare alleles given the observed data ($0 \leq A_i \leq 2$), the software mach2dat was used (`http://www.sph.umich.edu/csg/yli/mach/index.html`) (Li et al., 2009, 2010). For the hard call genotypes, where $A_i \in \{0, 1, 2\}$, we used the software PLINK version 1.07 (`http://pngu.mgh.harvard.edu/purcell/plink`) (Purcell et al., 2007). Figures were generated in the statistical software R version 2.9.0 (R Development Core Team, 2009).

We grouped the SNPs into four categories: SNPs genotyped on both chips; SNPs genotyped on Affy and imputed for the Illumina controls; SNPs genotyped on Illumina and imputed for the Affy controls; and SNPs imputed for both groups. The false positives found among SNPs genotyped on both platforms can be thought of as a baseline error rate against which to compare the other three groups. For each group of SNPs we summarized the error rates using two quantities: the Genomic Control λ and the percentage of SNPs with p-value less than 5×10^{-8}. For χ^2 test statistics X_i, $i = 1, \ldots, m$, the Genomic Control λ is defined as

$$\lambda = \underset{i=1,\ldots,m}{\text{median}}\{X_i\}/0.455$$

where 0.455 is approximately the theoretical median of a χ^2_1 distribution (Devlin and Roeder, 1999). Our model assumes the null distribution of each X_i is χ^2_1, so if this assumption is valid, we should have $\lambda \approx 1$. A value of $\lambda > 1$ suggests that the observed variance of the test statistic is larger than the theoretical variance, which will tend to increase the number of false positives. We also calculated the percentage of SNPs significant at the 5×10^{-8} significance level, a standard significance level used for GWA studies (McCarthy et al., 2008). Assuming the genotype is measured accurately, we don't expect genotype frequency differences between our cases and controls, because they are both samples of healthy women used as control groups for other studies. Thus, we should see very few SNPs with such significant p-values (approximately 1 out of every 20,000,000 independent tests).

When $\lambda > 1$ and the percentage of SNPs significant at the 5×10^{-8} level was more than expected in our null setting, we explored 3 methods for controlling for the error inflation:

1.2.1 Method 1

We investigated whether we could capture the platform effect using PCs. To do this, we used EIGENSTRAT (`http://genepath.med.harvard.edu/~reich/Software.htm`) (Patterson et al., 2006; Price et al., 2006). In a typical application of this program, the first few PCs are calculated and included as covariates in logistic regression to capture and control for population stratification. An example in Price et al. (2006) suggests the possibility of some components capturing lab and batch effects as well. We calculated the first ten PCs and assessed how well they correlated with platform effect. Then we attempted to include these components as covariates in logistic regression models predicting case-control status from each SNP. We did this in two ways: first, we calculated the PCs using all measured and imputed SNPs; second, we restricted to SNPs in each of the four categories, and calculated PCs using only those SNPs (e.g., using only SNPs measured on one chip and imputed in the other).

1.2.2 Method 2

When missing genotypes are imputed by MaCH, each SNP has an R^2 value associated with it that quantifies the quality of the imputation. The R^2 value is an estimate of the squared correlation between the imputed genotype and the actual genotype, so a higher R^2 corresponds to a SNP imputed with more certainty. Standard advice is to restrict to SNPs with $R^2 > 0.3$, which we did (Scott et al., 2007). It is expected that this will remove 70% of poorly imputed SNPs while keeping 99.5% of better imputed SNPs (Li et al., 2010). To reduce the error inflation in our less standard setting, we considered restricting to SNPs imputed at even higher quality.

Focusing on SNPs measured on one chip and imputed in the other, we considered removing SNPs with imputation $R^2 < 0.5, 0.75, 0.9, 0.95$ and 0.99. After thresholding by each value of R^2, we calculated λ and the percentage of SNPs with $p < 5 \times 10^{-8}$. We kept track of the number of SNPs still available for analysis at each threshold.

We also constructed an ROC curve to assess the discriminatory ability of this method. We labeled SNPs with $p < 5 \times 10^{-8}$ as "problematic." As we varied the R^2 threshold between 0 and 1, we compared how many problematic SNPs were being detected (sensitivity) to how many non-problematic SNPs were being excluded due to low R^2 ($1-$specificity).

1.2.3 Method 3

The genotype distributions for some SNPs may differ markedly across platforms due to genotyping artifact or differences in imputation quality. These differences may be identified even in relatively small samples. We explored the possibility of genotyping a small number of additional controls along with the cases, which could be used to identify and eliminate the problematic SNPs. Researchers would perform a preliminary analysis comparing the additional controls to the original controls, and any SNP significant in this preliminary analysis would be discarded. Researchers could then perform standard association tests between cases and controls using the remaining SNPs.

We randomly selected 1000 subjects from the 1,038 on Illumina to serve as controls, and 1000 subjects from the 1,672 on Affy to serve as cases. Then from the remaining 672 subjects on Affy, we selected n additional subjects to serve as controls genotyped alongside cases on the Affy platform. We first performed a screening step, in which we compared these n Affy controls to the 1000 Illumina controls and eliminated SNPs significant at level α. Then, restricting to SNPs that passed this screening, we performed the main analysis, comparing the 1000 Illumina controls to the 1000 Affy cases, and calculated the Genomic control λ and the percentage of SNPs with $p < 5 \times 10^{-8}$ in this main analysis. We did this calculation for $n = 100, 300$ and 500, and for $\alpha = 0.001, 0.01, 0.1$, and 0.2. We also constructed ROC curves to assess the discriminatory ability of this method while varying α, the screening threshold. That is, as we varied the α screening threshold between 0 and 1, we compared how many problematic SNPs (in the main analysis of 1000 Illumina controls vs. 1000 Affy cases) were being detected to how many non-problematic SNPs were being excluded.

1.3 Results

Figure 1.1 summarizes the results of a standard logistic regression analysis, where SNPs are grouped by MAF. For each collection of SNPs, we found the Genomic Control λ (in black) and the percentage of SNPs with $p < 5 \times 10^{-8}$ (in gray). Results from the soft call analysis are shown in solid lines, while those from the hard call analysis are shown in dashed lines. In Figure 1.1a, we see that $\lambda \approx 1$ among the 139,732 SNPs measured on both chips, and the percentage of highly significant SNPs is close to 0 across all MAFs; the error measures in this setting are virtually identical whether we use hard call or soft call imputation. Thus,

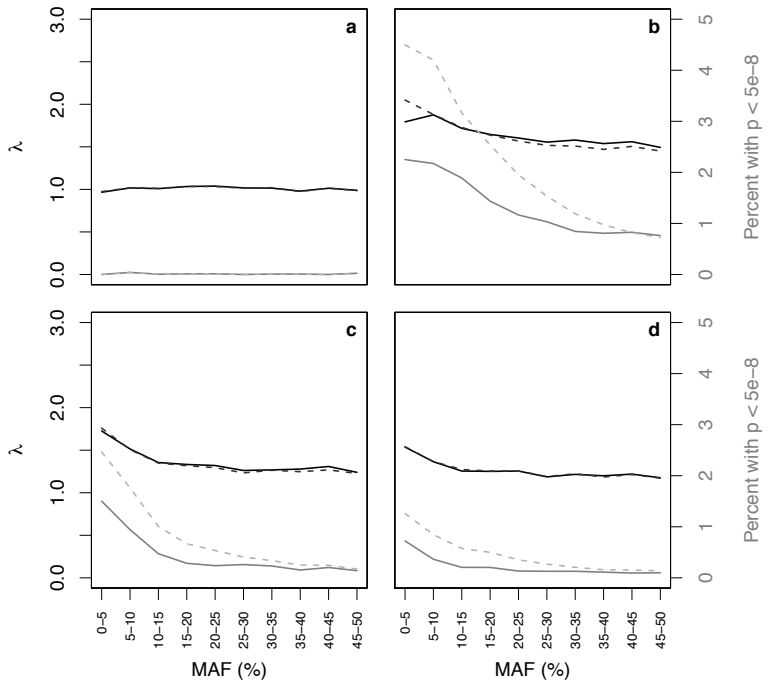

Figure 1.1: Black and gray lines represent λ values and the percentages of p-values less than 5×10^{-8}, respectively, for SNPs grouped by minor allele frequency (MAF) in four settings: **a** SNPs genotyped on both Affy and Illumina platforms; **b** SNPs genotyped on Illumina platform and imputed for the Affy controls; **c** SNPs genotyped on Affy platform and imputed for the Illumina controls; and **d** SNPs imputed for both groups. Solid lines are from soft call analysis and dashed lines are from hard call analysis. Note that in some places (particularly in panel **a**) the solid and dashed lines are indistinguishable because the results from the soft call and hard call analyses were very similar.

when we consider only the SNPs measured on both chips, we have no evidence from these two measures that the distribution of the test statistics deviates from the null.

However, among the 357,361 SNPs measured on Illumina and imputed on Affy (Figure 1.1b), we see an overall increase in λ to 1.6. We see an increase in the percentage of highly significant SNPs to 1.3% when using soft call genotypes, and to 2.1% when using hard call genotypes. Thus, when using hard call genotypes, 7,644 SNPs are being declared significant at the 5×10^{-8} level. These increases are most prominent among SNPs with low MAF, as shown in the Figure. The Type I error inflation is also apparent, though less dramatic, among the 458,034 SNPs measured on Affy and imputed on Illumina (Figure 1.1c) where $\lambda = 1.3$ overall, and where we are seeing 0.4% highly significant SNPs when using soft calls and 0.8% highly significant SNPs when using hard calls; we see similar numbers among the 1,392,682 SNPs imputed in both (Figure 1.1d). Results were largely unchanged when we first pooled the two groups and then imputed.

To try to correct these problems, we applied the three described methods. Here, we present results for the SNPs measured on Illumina and imputed on Affy for simplicity; results were similar in the other two problematic cases.

1.3.1 Method 1

We found the first ten PCs using hard call genotypes because those are currently supported by EIGENSTRAT. We did this once using all SNPs, and once restricting to SNPs measured on Illumina and imputed on Affy. Results were similar in the two approaches, and results from the latter are shown. The top three PCs are plotted against one another in Figure 1.2. We see that the second PC completely separates the cases (i.e., the Affy controls) and controls (the Illumina controls). Thus, when these PCs are included in a logistic regression predicting case-control status, we get a complete separation of data points, and the models cannot be fit.

1.3.2 Method 2

We considered restricting to SNPs imputed with increasingly higher quality, as quantified by the imputation R^2. Results for the soft call genotypes are shown in Table 1.1. As the R^2 threshold was increased, our summary measures improved; however, this happened at the expense of losing SNPs for analysis, which

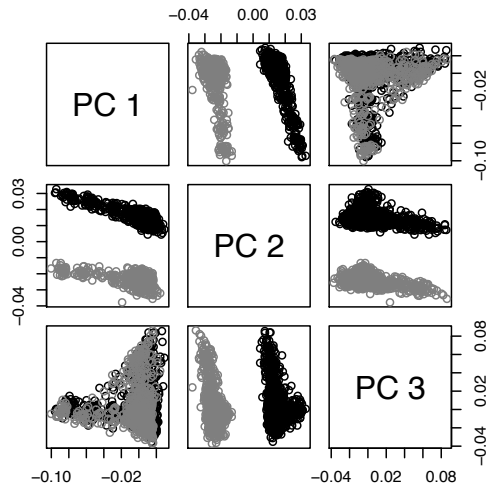

Figure 1.2: Top 3 principal components (PCs), among SNPs genotyped in the Illumina controls and imputed using hard calls in the Affy controls, plotted against one another. Affy samples are plotted in black; Illumina samples are plotted in gray.

Table 1.1: Among SNPs genotyped in the Illumina controls and imputed using soft calls among the Affy controls, values of λ and percentages of SNPs with $p < 5 \times 10^{-8}$ when we restrict to SNPs with imputation quality R^2 larger than the given thresholds, as detailed in Method 2. Also listed are the percentages of total SNPs remaining for analysis at each threshold.

	R^2 threshold					
	0.3	0.5	0.75	0.9	0.95	0.99
λ	1.6	1.6	1.4	1.2	1.1	1.04
% SNPs with $p < 5 \times 10^{-8}$	1.3	0.87	0.36	0.15	0.09	0.05
% SNPs remaining for analysis	100	97	87	71	59	31

reflects some loss of power. It should also be noted that even at the most stringent threshold listed, $R^2 > 0.99$, when we've excluded nearly 70% of the SNPs, there remain 57 SNPs with $p < 5 \times 10^{-8}$. Figure 1.3 shows the discriminatory ability of this method as we vary the R^2 threshold.

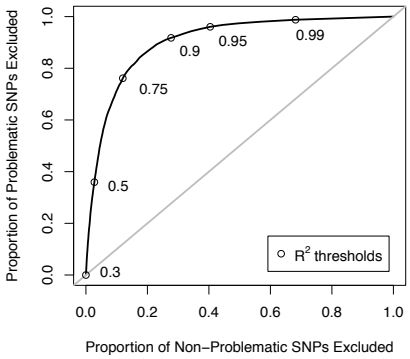

Figure 1.3: Among SNPs genotyped in the Illumina controls and imputed using soft calls among the Affy controls, discrimination of the R^2 criterion described in Method 2, as the R^2-threshold varies. The y-axis is the sensitivity, the proportion of highly significant SNPs which are excluded; the x-axis is $1-$specificity, the proportion of non-significant SNPs which are excluded. R^2 threshold choices between 0.3 and 0.99 are pointed out along the curve.

1.3.3 Method 3

For various thresholds ($\alpha = 0.001, 0.01, 0.1, 0.2$) and various numbers of additional controls on Affy ($n = 100, 300, 500$) we removed SNPs significant at level α in a preliminary analysis comparing the n additional Affy controls to the 1000 original Illumina controls. We then performed standard logistic regressions comparing the 1000 Illumina controls to the 1000 Affy cases using the remaining SNPs. The genomic control λ and the percentage of highly significant SNPs were calculated; results for the soft call genotypes are shown in Table 1.2. As n increased and α increased, our summary measures improved. This again happened at the expense of losing SNPs for analysis, but not as quickly as in Method 2. Figure 1.4 shows the discriminatory ability of this method for each n as we vary the α screening threshold.

Table 1.2: Results from the main analysis comparing 1000 Affy cases and 1000 Illumina controls, among SNPs remaining after a preliminary screen in which we compare $n = 100, 300,$ or 500 additional controls on Affy to 1000 controls on Illumina and remove SNPs significant at level α, as detailed in Method 3. Among SNPs genotyped in the Illumina controls and imputed using soft calls among the Affy controls, values of λ and percentages of SNPs with $p < 5 \times 10^{-8}$ are presented, along with the percentages of total SNPs remaining for analysis at each threshold.

		α threshold				
		0	0.001	0.01	0.1	0.2
100 additional controls	λ	1.5	1.5	1.5	1.4	1.3
	% SNPs with $p < 5 \times 10^{-8}$	0.83	0.59	0.38	0.11	0.07
	% SNPs remaining for analysis	100	100	98	88	78
300 additional controls	λ	1.5	1.5	1.4	1.2	1.2
	% SNPs with $p < 5 \times 10^{-8}$	0.83	0.14	0.04	0.01	0.002
	% SNPs remaining for analysis	100	99	96	85	74
500 additional controls	λ	1.5	1.4	1.4	1.2	1.1
	% SNPs with $p < 5 \times 10^{-8}$	0.83	0.03	0.01	0.001	0.0004
	% SNPs remaining for analysis	100	98	95	82	72

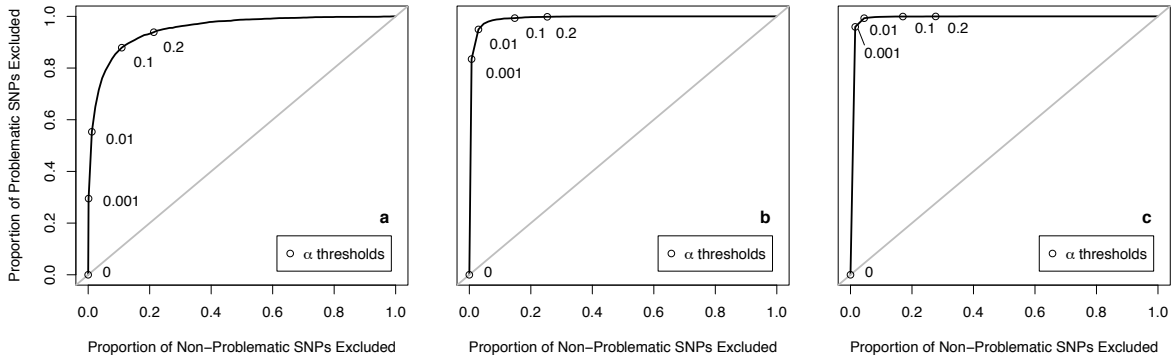

Figure 1.4: Among SNPs genotyped in the Illumina controls and imputed using soft calls among the Affy controls, discrimination of the preliminary screening criterion described in Method 3, as the α-threshold varies. The y-axis is the sensitivity, the proportion of highly significant SNPs which are excluded; the x-axis is $1-$specificity, the proportion of non-significant SNPs which are excluded. Plots shown are for **a** $n = 100$, **b** $n = 300$, and **c** $n = 500$ additional controls. α threshold choices between 0.001 and 0.2 are pointed out along the curves.

1.4 Discussion

We observed a large number of highly significant SNPs after imputation in a study comparing two healthy control groups genotyped on different platforms. Because both control groups are nested in the NHS and chosen using similar criteria, we expect no SNPs to significantly distinguish the two groups in the absence of measurement error, and we expect no differential population substructure. Thus, statistically significant SNPs are false positives, and must be due to genotyping or imputation error. Furthermore, because we see almost no inflation in Type I error among SNPs actually genotyped on both chips (Figure 1.1a), the false positives do not appear to result from genotyping error. Rather, the inflation in Type I error is seen among SNPs measured in one group and imputed in the other (and among SNPs imputed in both). In this setting, it would be detrimental to avoid imputation altogether since only about a quarter of the SNPs genotyped on each platform overlap, so that three-quarters of the SNPs on each chip would be unusable without any imputation. Thus, we need to understand the errors being introduced by imputation and attempt to control for them.

We believe that the inflation in Type I error is due to bias introduced by the differential imputation. The imputation uses individuals in the HapMap as a reference panel, and it seems plausible that estimates in the HapMap, particularly for rare alleles, may diverge from the allele frequencies observed in our population. Thus, if a rare allele has similar frequencies in our cases and controls but is not well covered in the HapMap, the p-value calculated when the SNP is measured in one group and imputed in the other will tend to be smaller than the p-value that would arise if that SNP were measured in both groups. Moreover, among SNPs with low MAF, Moskvina et al. (2006) showed that even modest differential errors in genotype calling can yield an inflation in Type I error. Generalized to our setting, this suggests that even slight differential errors in imputation among SNPs with low MAF would lead to false positive associations. This is borne out by our results, where we see larger numbers of highly significant p-values among SNPs with low MAF, as shown in Figure 1.1.

The percentage of highly significant SNPs is noticeably larger in the hard call analysis than in the soft call analysis. This is because the soft call imputations better account for uncertainty in the imputed values. We recommend using soft calls, or another technique that accounts for imputation uncertainty, in order to reduce false positives. It is worth considering whether we could somehow alter the imputation methods themselves to avoid these false positives altogether; however, it is unclear to what extent this is

possible. Imputation algorithms are limited by the information they are provided. For some platforms, the genotyped SNPs provide enough information to accurately infer an unobserved SNP; for other platforms, they do not, regardless of the imputation algorithm. Moreover, current imputation methods have good accuracy, particularly for SNPs with higher imputation R^2 (Li et al., 2010), yet even SNPs with high R^2 appear among our false positives. This suggests that even well-imputed SNPs can be falsely significant when the imputation error is differential.

The inflation in Type I error appears to be most dramatic among SNPs measured in Illumina and imputed in Affy. We suspect that this is because Illumina uses HapMap for SNP selection, and we used HapMap for SNP imputation. When we considered SNPs common to both chips, the distribution of test statistics was what we expect under the null, suggesting that the actual genotyping across the two chips is in good agreement.

When we attempted to reduce the error inflation using PCs, in Method 1, we observed a complete separation of the two control groups. This complete separation shows the difficulty of controlling for platform effect by simply adjusting for PCs. Including the PCs as covariates in the model is equivalent to including case-control status as a covariate, and thus there does not appear to be a direct way to use those PCs to resolve the error inflation problem. Furthermore, any method using the PCs would likely wash out all differences between cases and controls in a non-null setting. Thus, it makes sense to focus on approaches that filter out problematic SNPs and exclude them from subsequent analysis. Methods 2 and 3 are two such approaches.

In Method 2, we used imputation quality to filter SNPs before performing any association tests. This approach improved the results and does not require genotyping any additional controls. It reduces the number of SNPs available for analysis, but still allows the use of more SNPs than just those actually genotyped on both platforms. However, in our example of SNPs genotyped on Illumina and imputed on Affy, even after filtering to SNPs imputed with $R^2 > 0.99$ (allowing us to retain only 30% of SNPs), we are left with 57 SNPs with highly significant p-values out of 112,249 remaining SNPs. So if this method is used, researchers should be prepared to sift through many false positives in a second stage analysis to find any true associations. Furthermore, this method will tend to reduce power to detect SNPs in regions with low linkage disequilibrium. Beecham et al. (2010) demonstrated this problem by pooling two case-control GWA studies for Alzheimer disease which had been genotyped on different chips, and testing for associations in the *APOE* gene, which is known to be strongly associated with risk. They used imputation to produce com-

mensurable data sets, and filtered out SNPs according to imputation quality. They found that even though each study separately found strong associations in the *APOE* gene, there was no association in the pooled analysis, because many SNPs had been excluded due to low imputation quality measures caused by weak linkage disequilibrium in the region.

In Method 3, we propose genotyping a small number of additional controls alongside the cases and performing a preliminary step of filtering SNPs by comparing these additional controls to the original controls. This approach also improves results, but at increased monetary cost. It should, however, retain more non-artifactual SNPs while reducing the number of artifactual SNPs. In our example of 1000 cases and 1000 controls, it appeared that genotyping 300 additional controls alongside cases would allow researchers to filter out most of the false positives — with $\alpha = 0.2$, only 5 highly significant SNPs were left among the SNPs genotyped on Illumina and imputed on Affy, with 264,519 (74%) remaining for analysis. We believe these results would be the same if we had new cases and controls on Illumina and a separate control group on Affy — we merely consider this setting because it made best use of the subjects available on each chip. This method is in line with the discussion in McCarthy et al. (2008) regarding the use of historical controls. McCarthy et al. listed many possible sources of systematic error that might arise in the use of historical controls, and recommended always genotyping some ethnically matched controls alongside cases on the same platform.

It may also be worth considering a related study design in which very little error inflation was seen, which was considered by Howie et al. (2009). In their setting, a central control group in the WTCCC was genotyped on both Affy and Illumina, while different case groups from different disease studies were genotyped on just one of these platforms. The authors were interested in whether imputing SNPs missing in cases using both the HapMap and the central control group as a reference panel led to inflated Type I error. To assess this, they compared the central control group with another control group genotyped on Affy alone. They imputed SNPs missing in this new control group and then performed association tests. They found very few significant results, which demonstrated minimal inflation of Type I error in this setting. Their methods differ slightly from ours; however, we believe that the most important difference was the nested structure of their design – that is, that their central control group had SNPs from both Affy and Illumina chips, rather than Illumina alone. A comparison of their results and ours suggests that if a central control group is going to be reused for different diseases, it may be wise to invest in genotyping the central control group on multiple platforms. A similar conclusion is offered by Marchini and Howie (2010).

Researchers can make use of accumulating genetic resources to more economically and more powerfully understand the effects of genes on complex diseases. However, our findings add to a familiar refrain about GWA studies – that every step must be done with extreme care to avoid spurious results (McCarthy et al., 2008). More work needs to be done to determine the best approaches for combining cases and controls obtained from different sources. In any case-control study, cases and controls should be comparable, and recent studies have discussed how to control for differential population substructure when using publicly available controls (Zhuang et al., 2010; Luca et al., 2008). Our work emphasizes the need to control for technical errors caused by integrating data from different chips. Researchers attempting to use the sort of data we describe, in which cases and controls are genotyped on different chips, need to be aware of the high potential for false positives after imputation, and must guard against it or control for it. In particular, it is vitally important to technically validate any SNPs that appear significant before reporting them, by regenotyping those SNPs on an independent platform – considered best practice in any GWA study, it is all the more important here where the chance of false positive results due to differential imputation is so high.

Omnibus Risk Assessment via Accelerated Failure Time Kernel Machine Modeling

Jennifer A. Sinnott and Tianxi Cai

Department of Biostatistics

Harvard School of Public Health

Abstract

Integrating genomic information with traditional clinical risk factors to improve the prediction of disease outcomes could profoundly change the practice of medicine. However, the large number of potential markers and possible complexity of the relationship between markers and disease make it difficult to construct accurate risk prediction models. Standard approaches for identifying important markers often rely on marginal associations and may not capture non-linear or interactive effects. In recent years, much work has been done to group genes into pathways and networks. Integrating such biological knowledge into statistical learning could potentially improve model interpretability and reliability. One effective approach is to employ a kernel machine (KM) framework, which has been used to make predictions for various types of outcomes (Scholkopf and Smola, 2002; Liu et al., 2007, 2008). For survival outcomes, regression modeling and testing procedures have been derived under a proportional hazards (PH) assumption (Li and Luan, 2003; Cai et al., 2011). In this paper, we propose KM regression under the accelerated failure time model, a useful alternative to the PH model. We derive a pseudo-score statistic for testing and a risk score for prediction of survival. To approximate the null distribution of our test statistic, we propose resampling procedures which also enable us to develop alternative robust testing procedures that combine information across kernels. Numerical studies show that the testing and prediction procedures perform well. The methods are illustrated with an application in breast cancer.

2.1 Introduction

Understanding the relationship between genomic markers and complex disease could have a profound impact on biological research, pharmacology, and medicine. Many traditional approaches for quantifying this relationship identify individual markers with marginal associations with disease; however, resulting lists of differentially expressed genes can be hard to interpret or replicate, and may not include truly important markers with modest, nonlinear, or interactive effects. An appealing alternative is to leverage current biological knowledge by grouping markers into networks and pathways consisting of genes thought to work together. Working at the pathway level reduces dimensionality, which decreases the number of statistical hypotheses to be tested and can improve power to detect associations. Moreover, lists of important pathways can be easier to interpret than lists of genes because pathways are often defined by known or hypothesized functions, thus facilitating the generation of mechanistic hypotheses and the identification of potential avenues for intervention. Ideally, pathway methods should identify genes which may individually have only modest associations with outcome but which together have a substantial joint effect. They should also allow for complicated relationships among genes in the pathway that could reflect more complex biological signals.

Kernel machine (KM) methods (Scholkopf and Smola, 2002) are attractive tools for relating biological pathways to disease outcomes because they can capture complex effects without explicit specification of the form of those effects, and because they can leverage the within-pathway correlation which is likely to exist in genomic data. For non-censored outcomes, KM regression and testing procedures have been proposed in Liu et al. (2007) and Liu et al. (2008). For survival outcomes, Li and Luan (2003) and Cai et al. (2011) proposed KM testing and estimation procedures under the proportional hazards (PH) model. However, when the PH assumption fails to hold, these procedures may have little power to identify important pathways or accurately predict risk. In this paper, we propose KM methods for survival outcomes under the accelerated failure time (AFT) model (Kalbfleisch et al., 1980), a useful alternative to the PH model.

The standard semiparametric AFT model relates covariates to log-survival time through a linear model. This model is appealingly interpretable, but has been used less than the Cox model in part because it can be somewhat challenging to fit in the presence of censoring. Inference procedures for the regression parameters under the AFT model include the inverse probability weighting (IPW), Buckley-James, rank-based, and sieve likelihood (SL) methods (Buckley and James, 1979; Koul et al., 1981; Tsiatis,

1990; Wei, 1992; Zeng and Lin, 2007). The IPW approach requires that the conditional censoring distribution be correctly specified and that the support of the censoring contain that of the failure time, which are both unlikely in practice. The Buckley-James procedure relies on the identifiability of the entire residual distribution, which may not be available in the presence of censoring. The SL estimator is fully efficient, but could be computationally challenging because it requires estimating a non-parametric functional. The rank-based approach (Tsiatis, 1990; Ritov, 1990), which fits the model using a weighted log rank (WLR) estimating equation for various weights, has advantages including consistency of estimation without additional censoring assumptions, and an effective implementation developed in Jin et al. (2003) for making inference using resampling.

We propose the use of the AFT KM model to capture potentially complex, non-linear pathway effects on survival. We first develop a pseudo-score test under the KM framework using the WLR estimating function, using resampling procedures for inference. The WLR weight affects the estimation efficiency, and the kernel determines the structure of the pathway effect, but in practice it is often unclear which weight or kernel is most appropriate for a given dataset. To overcome this challenge, we propose omnibus testing procedures that pool information across weights and kernels. We demonstrate the effectiveness of this approach in simulations, where the power lost by using an omnibus test is minimal, but the power gained in some circumstances can be substantial.

When a pathway is associated with outcome, we may hope to use the pathway information to improve risk prediction. Under the AFT KM framework, we propose procedures for estimating the pathway effect in order to construct individual risk scores. Our simulation results suggest that these outperform those derived from standard linear effects models when the underlying effects are non-linear, while maintaining similar accuracy when the effects are linear. Recently, Liu et al. (2010) proposed the usage of the AFT KM model and presented a weighted least squares estimator. However, in the presence of censoring, it is unclear whether this estimator is consistent even under linear effects or how they calculate their criterion for tuning and model assessment, the relative root mean squared error. Because our procedure is derived using the WLR class of estimating equations, we are guaranteed consistent estimates of the underlying effects when the AFT model is correctly specified.

The rest of this paper is organized as follows. In Section 2.2 we introduce the AFT KM model. In Section 2.3 we present testing and estimation procedures using Gehan weights in the WLR. In Section 2.4 we present testing and estimation for general weights, as well as our omnibus testing procedure. Simulations

are presented in section 2.5 and our method is illustrated in section 2.6 in application to breast cancer data. Final remarks are in Section 2.7.

2.2 The Kernel Machine Accelerated Failure Time Model

Let T denote the survival time, $\mathbf{Z}$ be a $P \times 1$ vector of genetic measurements in a gene set, and $\mathbf{D}$ be a vector of clinical covariates such as age and gender. Due to censoring of T, we observe $X = \min\{T, C\}$ and $\Delta = I[T \leq C]$, where C is a censoring time that is assumed to be independent of T given $\mathbf{W} = (\mathbf{D}^\intercal, \mathbf{Z}^\intercal)^\intercal$. The observed data consist of n independent and identically distributed (iid) random vectors, $\mathcal{O} = \{(X_i, \Delta_i, \mathbf{W}_i) : i = 1, \ldots, n\}$.

To derive a prediction model for T based on $\mathbf{W}$, we consider the KM generalization of the AFT model:

$$\log T_i = \gamma^\intercal \mathbf{D}_i + h(\mathbf{Z}_i) + E_i,\ i = 1, \ldots, n. \tag{2.1}$$

where γ is the unknown covariate effect of $\mathbf{D}_i$, E_i is an iid error term independent of $\mathbf{W}_i$ with completely unspecified distribution, and $h(\cdot)$ is an unknown smooth function that belongs to $\mathcal{H}_K$, the Hilbert space generated by a given positive definite kernel $K(\cdot, \cdot; \rho)$. The kernel is a measure of similarity between two vectors of genetic measurements, and may depend on a possibly unknown scaling parameter ρ. Different choices of kernel K will yield different collections of possible functions $h(\cdot)$. For example, the *linear kernel* $K(\mathbf{z}_1, \mathbf{z}_2) = \mathbf{z}_1^\intercal \mathbf{z}_2$ leads to $h(\mathbf{z}) = \beta^\intercal \mathbf{z}$, a linear function of the covariates. The *quadratic kernel* $K(\mathbf{z}_1, \mathbf{z}_2; \rho) = (\rho + \mathbf{z}_1^\intercal \mathbf{z}_2)^2$ yields a Hilbert space $\mathcal{H}_K$ spanned by basis functions $\{z_j, z_j z_{j'} : j, j' = 1, \ldots, p\}$, which incorporates main effects, quadratic effects, and 2-way interactions. To allow for more complex nonlinear effects, one may consider the *Gaussian kernel*, defined by $K(\mathbf{z}_1, \mathbf{z}_2; \rho) = \exp\{-\|\mathbf{z}_1 - \mathbf{z}_2\|^2/\rho\}$. The resulting function space $\mathcal{H}_K$ is generated by the radial basis functions.

We will be interested in testing the null hypothesis $H_0 : h(\cdot) \equiv 0$, which suggests that the genes in the pathway do not affect survival given the covariates $\mathbf{D}$. Under the null, model (2.1) reduces to the standard AFT model for the clinical covariates,

$$\log T_i = \gamma^\intercal \mathbf{D}_i + E_i,\ i = 1, \ldots, n. \tag{2.2}$$

As described in Jin et al. (2003), the regression parameter γ can be estimated by solving the WLR estimating function,

$$\mathbf{U}_\phi(\gamma) = n^{-1} \sum_{i=1}^{n} \phi(\gamma, e_i(0; \gamma)) \Delta_i [\mathbf{D}_i - \overline{\mathbf{D}}(\gamma, e_i(0; \gamma))] \tag{2.3}$$

where $e_i(0;\gamma)$ is the residual $e_i(h;\gamma) = \log X_i - \gamma^\intercal \mathbf{D}_i - h(\mathbf{Z}_i)$ evaluated at $h = 0$, $\phi(\gamma, t)$ is a weight function, and $\overline{\mathbf{D}}(\gamma, t) = S^{(1)}(\gamma, t)/S^{(0)}(\gamma, t)$ for $S^{(k)}(\gamma, t) = n^{-1}\sum_{j=1}^n \mathbf{I}\{e_j(0;\gamma) \geq t\}\mathbf{D}_j^{\otimes k}$. Here, we use the notation $\boldsymbol{a}^{\otimes 0} = 1$, $\boldsymbol{a}^{\otimes 1} = \boldsymbol{a}$, and $\boldsymbol{a}^{\otimes 2} = \boldsymbol{a}\boldsymbol{a}^\intercal$ for any vector $\boldsymbol{a}$. $\mathbf{U}_\phi$ is not componentwise monotonic in γ in general, but it is when we use the Gehan weights $\phi = S^{(0)}$, which yield an estimating function $\mathbf{U}_G(\gamma) = n^{-2}\sum_{i=1}^n \sum_{j=1}^n \Delta_i[\mathbf{D}_i - \mathbf{D}_j]\mathbf{I}\{e_j(0;\gamma) \geq e_i(0;\gamma)\}$ which is the gradient of the convex Gehan objective function

$$L_G(\gamma) = n^{-2}\sum_{i=1}^n \sum_{j=1}^n \Delta_i |e_j(0;\gamma) - e_i(0;\gamma)|_+, \tag{2.4}$$

where $|a|_+ = a\mathbf{I}\{a > 0\}$. $L_G(\gamma)$ can be minimized to get an estimator $\widetilde{\gamma}_G$ which is consistent for γ in (2.2); Jin et al. (2003) also provides an iterative procedure to find estimators $\widetilde{\gamma}_\phi$ for general weights ϕ, which may be desirable because the variance of $\widetilde{\gamma}_\phi$ depends on ϕ.

When we do not assume the pathway effect $h(\cdot)$ is identically 0 in model (2.1), we may obtain estimators for γ and h by minimizing the penalized Gehan objective function analogous to equation (2.4),

$$L_G^R(\gamma, h) = n^{-2}\sum_{i=1}^n \sum_{j=1}^n \Delta_i |e_j(h;\gamma) - e_i(h;\gamma)|_+ + \frac{c^2}{2}\|h\|_{\mathcal{H}_K}^2, \tag{2.5}$$

where $\|h\|_{\mathcal{H}_K}$ is the norm of h in $\mathcal{H}_K$. By the representer theorem (Kimeldorf and Wahba, 1970), the minimizer of $L_G^R(\gamma, h)$ for h must take a dual form, $\widehat{h}(\mathbf{z}) = \sum_{l=1}^n \alpha_l K(\mathbf{Z}_l, \mathbf{z})$, where the α_l are unknown parameters. Here and in the sequel, we suppress ρ from K for ease of presentation, but note that testing and estimation may depend on ρ and that tuning ρ will be discussed when needed.

Using the dual representation, minimization of (2.5) is equivalent to minimization of

$$L_G^R(\alpha;\gamma) = n^{-2}\sum_{i=1}^n \sum_{j=1}^n \Delta_i |e_j(\alpha;\gamma) - e_i(\alpha;\gamma)|_+ + \frac{c^2}{2}\alpha^\intercal \mathbb{K}\alpha \tag{2.6}$$

where $\mathbb{K}$ is the kernel matrix, whose $(i,j)^{\text{th}}$ entry is $K(\mathbf{Z}_i, \mathbf{Z}_j)$, and, with a slight abuse of notation, $e_i(\alpha;\gamma) = \log X_i - \gamma^\intercal \mathbf{D}_i - \alpha^\intercal \mathbf{K}_i$, for $\mathbf{K}_i$ the i^{th} row of $\mathbb{K}$. Hence, a plug-in estimator for h may be obtained by minimizing $L_G^R(\alpha;\gamma)$ with respect to α and γ. Moreover, we may view (2.6) as the Gehan objective function arising from the random effects AFT model

$$\log T_i = \gamma^\intercal \mathbf{D}_i + \alpha^\intercal \mathbf{K}_i + E_i, \quad \alpha = \tau\epsilon, \quad E(\epsilon) = 0, \quad \text{var}(\epsilon) = \mathbb{K}^-, \tag{2.7}$$

with ϵ multivariate normal and $c^{-1} = \tau$. Here $\mathbb{K}^-$ is the Moore-Penrose generalized inverse of $\mathbb{K}$. Analogous connections between penalized KM models and the mixed model framework were successfully used to fit KM regression in other models (Liu et al., 2007, 2008; Cai et al., 2011).

2.3 Testing and Estimation with Gehan Weights

2.3.1 The Pseudo-Score Statistic

To identify pathways associated with survival, we propose the use of the AFT KM framework and derive testing procedures for $H_0 : h(\cdot) = 0$ in model (2.1). Here we derive a KM pseudo-score test of H_0 using the WLR estimating function with Gehan weights; in Section 2.4 we show how to extend the test to more general weights.

By using the mixed effects formulation (2.7) as a working model, we see that the hypothesis $H_0 : h(\cdot) = 0$ is equivalent to testing $H_0 : \tau = 0$, and so we can derive a KM pseudo-score test procedure by writing the penalized Gehan objective function (2.6) as a function of τ conditional on random effects ϵ, $L_{G,\epsilon}(\tau;\gamma)+\epsilon^{\intercal}\mathbb{K}\epsilon$, where

$$L_{G,\epsilon}(\tau;\gamma) = n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}\Delta_i|e_j(\tau\epsilon;\gamma) - e_i(\tau\epsilon;\gamma)|_+. \tag{2.8}$$

If γ is known, a pseudo-score statistic can be obtained as

$$\widehat{Q}(\gamma) = E_\epsilon\left[\left\{\nabla_\tau L_{G,\epsilon}(\tau;\gamma)|_{\tau=0}\right\}^2\middle|\,\mathcal{O}\right],$$

where ∇_τ is the partial derivative with respect to τ. When γ is unknown, as would be the case in practice, we estimate $\widetilde{\gamma}_G$ under the null as described in Section 2.2, and define our pseudo-score statistic as $\widehat{Q} = \widehat{Q}(\widetilde{\gamma}_G)$. Since $\nabla_\tau L_{G,\epsilon}(\tau;\gamma)|_{\tau=0} = \epsilon^{\intercal}\mathbb{K}\widehat{\mathbf{R}}(\gamma)$, where $\widehat{\mathbf{R}}(\gamma) = (\widehat{R}_1(\gamma),\ldots,\widehat{R}_n(\gamma))^{\intercal}$ and $\widehat{R}_k(\gamma) = n^{-2}\sum_{j=1}^{n}\left[\Delta_k\mathbf{I}\{e_j(0;\gamma)\geq e_k(0;\gamma)\} - \Delta_j\mathbf{I}\{e_k(0;\gamma)\geq e_j(0;\gamma)\}\right]$, we can see that $\widehat{Q} = \widehat{\mathbf{R}}(\gamma)^{\intercal}\mathbb{K}\widehat{\mathbf{R}}(\gamma)$. We may further rewrite this by employing a spectral decomposition for $\mathbb{K}$. If we let the eigenvalues and associated eigenvectors of $\mathbb{K}$ be $\widehat{\lambda}_l$ and $\widehat{\boldsymbol{\zeta}}_l$ respectively, for $l = 1,\ldots,n$, where we assume that $\widehat{\lambda}_1 \geq \cdots \geq \widehat{\lambda}_n$ and that the $\widehat{\boldsymbol{\zeta}}_l$ have norm 1, then we may write $\mathbb{K} = \widetilde{\mathbb{B}}_n\widetilde{\mathbb{B}}_n^{\intercal}$, where $\widetilde{\mathbb{B}}_n = \left(\sqrt{\widehat{\lambda}_1}\widehat{\zeta}_1\cdots\sqrt{\widehat{\lambda}_n}\widehat{\zeta}_n\right)$. This allows us to reexpress $\widehat{Q}(\gamma)$ as:

$$\widehat{Q}(\gamma) = \widehat{\mathbf{U}}_G(\gamma)^{\intercal}\widehat{\mathbf{U}}_G(\gamma)$$

where

$$\widehat{\mathbf{U}}_G(\gamma) = \widetilde{\mathbb{B}}_n^{\intercal}\widehat{\mathbf{R}}(\gamma) = n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}\Delta_i(\widetilde{\mathbf{B}}_{ni} - \widetilde{\mathbf{B}}_{nj})\mathbf{I}\{e_j(0;\gamma)\geq e_i(0;\gamma)\}. \tag{2.9}$$

Here $\widetilde{\mathbf{B}}_{ni}$ is the i^{th} row of $\widetilde{\mathbb{B}}_n$. We note that the form of $\widehat{\mathbf{U}}_G$ mimics that of the standard Gehan WLR function $\mathbf{U}_G$ defined in Section 2.2, with the $\widetilde{\mathbf{B}}_{ni}$ taking the role of the covariates $\mathbf{D}_i$.

2.3.2 Approximating the Null Distribution of the Score Statistic

In the Web Appendix, we outline the derivation of the asymptotic distribution $\mathcal{Q}$ of $n\widehat{Q}$. This distribution generally does not have an explicit form, so to approximate the null distribution of $\widehat{Q}$ in finite samples, we propose a perturbation approach similar to that used in Jin et al. (2003) for the linear effects AFT model.

Let $\mathcal{V} = (\mathcal{V}_1, \ldots, \mathcal{V}_n)$, and let the $\{\mathcal{V}_i\}$ be iid random variables with mean 1 and variance 1. We first find $\widetilde{\gamma}_G^*$, the minimizer of $L_G^*(\gamma) = n^{-2}\sum_{i=1}^n \sum_{j=1}^n \Delta_i |e_j(0;\gamma) - e_i(0;\gamma)|_+ \mathcal{V}_i\mathcal{V}_j$. Then, we calculate the perturbation $\widehat{\mathbf{U}}_G^*(\widetilde{\gamma}_G^*)$ of $\widehat{\mathbf{U}}_G(\widetilde{\gamma}_G)$ as:

$$\widehat{\mathbf{U}}_G^*(\widetilde{\gamma}_G^*) = n^{-2}\sum_{i=1}^n\sum_{j=1}^n \Delta_i(\widetilde{\mathbf{B}}_{ni} - \widetilde{\mathbf{B}}_{nj})\mathbf{I}\{e_j(0;\widetilde{\gamma}_G^*) \geq e_i(0;\widetilde{\gamma}_G^*)\}\mathcal{V}_i\mathcal{V}_j. \tag{2.10}$$

Using similar arguments as Jin et al. (2001), we can show that under H_0, the distribution of $n^{\frac{1}{2}}\widehat{\mathbf{U}}_G(\widetilde{\gamma}_G)$ can be approximated by the distribution of $n^{\frac{1}{2}}\{\widehat{\mathbf{U}}_G^*(\widetilde{\gamma}_G^*) - \widehat{\mathbf{U}}_G(\widetilde{\gamma}_G)\}$ conditional on $\mathcal{O}$. Thus, the null distribution of $\widehat{Q}$ can be approximated by the distribution of $\widehat{Q}^* = \widehat{Q}^*(\widetilde{\gamma}_G^*) = \{\widehat{\mathbf{U}}_G^*(\widetilde{\gamma}_G^*) - \widehat{\mathbf{U}}_G(\widetilde{\gamma}_G)\}^\intercal\{\widehat{\mathbf{U}}_G^*(\widetilde{\gamma}_G^*) - \widehat{\mathbf{U}}_G(\widetilde{\gamma}_G)\}$ given $\mathcal{O}$.

To calculate a p-value for testing H_0, we may generate a large number of realizations of $\mathcal{V}$, say $\mathcal{V}_{(1)}, \ldots, \mathcal{V}_{(B)}$, and use them to generate $\widehat{Q}_{(1)}^*, \ldots, \widehat{Q}_{(B)}^*$ Then the p-value of the test can be obtained by $\widehat{p} = \#\{\widehat{Q}_{(b)}^* \geq \widehat{Q}\}/B$. Alternatively, we may use the Satterthwaite method to approximate the null distribution using a scaled χ^2 distribution, $c_0\chi^2_{d_0}$. We estimate c_0 and d_0 by matching moments with the estimated null $\{\widehat{Q}_{(b)}^*\}$, and calculate $\widehat{p}_{\chi^2} = 1 - F(\widehat{Q})$, where F is the distribution function of a $\widehat{c}_0\chi^2_{\widehat{d}_0}$ random variable. Similar approximations have performed well for other models (Liu et al., 2007, 2008; Cai et al., 2011).

For the linear kernel, which does not rely on an additional parameter ρ, we may use the p-values $\widehat{p}_{\text{obs}} = \widehat{p}$ or $\widehat{p}_{\chi^2,\text{obs}} = \widehat{p}_{\chi^2}$ directly. For the quadratic and Gaussian kernels, the test statistic $\widehat{Q}_{\text{obs}}$ may depend on the parameter ρ. Since the kernel matrix $\mathbb{K}(\rho)$ drops out of the model under the null, the parameter ρ is not estimable (Davies, 1987). Instead, we propose to use as a test statistic:

$$\widehat{T}_{\mathcal{I}} = \inf_{\rho\in\mathcal{I}}\left\{\widehat{p}_{\chi^2}(\rho)\right\}, \tag{2.11}$$

where $\widehat{p}_{\chi^2}(\rho)$ denotes the p-value from the χ^2 approximation derived under kernel function $K(\cdot,\cdot;\rho)$ and $\mathcal{I}$ is an appropriately chosen range for ρ. The final p-value then is $\widehat{p}_{\text{obs}} = \#\{\widehat{T}^*_{\mathcal{I},(b)} \leq \widehat{T}_{\mathcal{I},\text{obs}}\}/B$. For the Gaussian and quadratic kernels, we determine the range $\mathcal{I}$ of ρ by requiring that the associated kernel matrices $\mathbb{K}(\rho)$ have eigenvalues $\widehat{\lambda}_i$ that decay at a polynomial rate $O(i^{-\alpha})$ for some range of $\alpha > 1$. This approach is motivated by work in Braun (2005) which bounds the error due to projecting the feature space

onto the first r principal components using terms whose behavior depends on the decay properties of the eigenvalues. By experimentation we found that $\alpha \in [1.75, 6]$ yielded a reasonable range of feature space complexity. To implement this, for each ρ, we regressed the logarithms of the top 95% of eigenvalues on their indices and used the regression coefficient as an estimate of α; we considered values ρ whose associated α fell in the specified range.

2.3.3 Estimation and the Kernel Principal Components Analysis Approximation

To construct a risk score $\gamma^{\intercal}\mathbf{D} + h(\mathbf{Z})$ for predicting T, we can minimize the penalized Gehan objective function (2.6) to get estimates for γ and α. The parameter c in (2.6) controls the smoothness of the resulting estimator for h, and can be chosen by cross-validation to minimize, for example, the unpenalized Gehan objective function. For kernels that depend on a tuning parameter ρ, we use the value of ρ that minimized $\widehat{T}_{\mathcal{I}}$ in (2.11). This choice is intuitively appealing because this ρ corresponds to the kernel $K(\cdot, \cdot; \rho)$ that produced the most evidence that $h(\cdot) \neq 0$.

While estimating (γ, α) in this manner is possible, it can be extremely computationally demanding because the dual parameter vector α has n components. Computation time can be decreased with little loss of accuracy by using a kernel PCA approximation for dimension reduction (Schölkopf et al., 1998; Mika et al., 1999). To do this, we again employ the spectral decomposition of $\mathbb{K}$, but retain only the first r principal components, where r is the smallest number for which $\sum_{i=1}^{r} \widehat{\lambda}_i / \sum_{i=1}^{n} \widehat{\lambda}_i \geq \mathfrak{p}$, where $\mathfrak{p}$ is some prespecified fraction. In many situations, r is significantly smaller than n. Ideally, the included eigenvectors encode aspects of maximal variability in the data, while the excluded eigenvectors capture noise.

Writing $\widetilde{\mathbb{B}}_r = \left(\sqrt{\widehat{\lambda}_1}\widehat{\zeta}_1 \cdots \sqrt{\widehat{\lambda}_r}\widehat{\zeta}_r\right)$ and $\widetilde{\mathbb{K}}_r = \widetilde{\mathbb{B}}_r \widetilde{\mathbb{B}}_r^{\intercal}$, we may consider an approximate working model analogous to (2.7) with $\widetilde{\mathbb{K}}_r$ in place of $\mathbb{K}$. In fact, our pseudo-score test may be derived exactly as above with $\widetilde{\mathbb{K}}_r$ and $\widetilde{\mathbb{B}}_r$ in place of $\mathbb{K}$ and $\widetilde{\mathbb{B}}_n$ to yield a slightly faster test procedure, which is what we use in practice. For estimation, we may apply the variable transformation $\beta = \widetilde{\mathbb{B}}_r^{\intercal}\alpha$ and rewrite the working model as:

$$\log \mathbf{T} = \mathbf{D}\gamma + \widetilde{\mathbb{B}}_r\beta + \mathbf{E}, \quad E(\beta) = 0, \quad \text{Var}(\beta) = \tau^2 \mathbb{I}_{r\times r}. \tag{2.12}$$

Then, instead of estimating (γ, α), we estimate (γ, β) by minimizing

$$L_G^R(\beta, \gamma) = n^{-2} \sum_{i=1}^{n}\sum_{j=1}^{n} \Delta_i |\widetilde{e}_j(\beta;\gamma) - \widetilde{e}_i(\beta;\gamma)|_+ + \frac{c^2}{2}\beta^{\intercal}\beta. \tag{2.13}$$

where $\widetilde{e}_i(\beta;\gamma) = \log X_i - \gamma^{\intercal}\mathbf{D}_i - \beta^{\intercal}\widetilde{\mathbf{B}}_{ri}$, for $\widetilde{\mathbf{B}}_{ri}$ the i^{th} row of $\widetilde{\mathbb{B}}_r$. The parameter β has r components, so

when the sample size is not small and the eigenvalues decay quickly, kernel PCA has the computational advantage of greatly reducing the number of unknown parameters to be estimated.

Formulating the estimation of h in this way is also appealing because it relates to the primal representation of h in the Hilbert space $\mathcal{H}_K$. The space $\mathcal{H}_K$ has a (possibly infinite) basis $\{\sqrt{\lambda_l}\zeta_l(\cdot), l = 1, 2, \ldots\}$ made up of eigenfunctions and eigenvalues $\zeta_l(\cdot)$ and λ_l of the integral transform $T : \mathcal{H}_K \to \mathcal{H}_K$ defined by $[Tf](\mathbf{z}) = \int K(\mathbf{z}, \mathbf{z}')f(\mathbf{z}')d\mathbf{z}'$, where $\lambda_1 \geq \lambda_2 \geq \cdots$. Any function $h \in \mathcal{H}_K$ may be written in its primal representation, $h(\mathbf{z}) = \sum_{l=1}^{\infty} \beta_l \sqrt{\lambda_l}\zeta_l(\mathbf{z})$, and if the eigenvalues decay quickly, can be well-approximated by a truncated sum $h_{r_0}(\mathbf{z}) = \sum_{l=1}^{r_0} \beta_l \sqrt{\lambda_l}\zeta_l(\mathbf{z})$ for some sufficiently large r_0. It has been shown that the eigenvalues and eigenvectors obtained based on $\mathbb{K}$ can be used to consistently estimate the underlying true eigenvalues and eigenfunctions (Koltchinskii and Giné, 2000; Braun, 2005). Thus, kernel PCA uses the eigendecomposition of $\mathbb{K}$ to approximate the basis of $\mathcal{H}_K$, and uses that to estimate h in its approximate primal form.

To apply the risk score to a future subject with predictors $\mathbf{W}_0 = (\mathbf{D}_0^\intercal, \mathbf{Z}_0^\intercal)^\intercal$, we may use the Nyström approximation method (Rasmussen and Williams, 2006) which relates these predictors to those in the training data. Letting $(\widehat{\gamma}^\intercal, \widehat{\beta}^\intercal)$ be the estimators for $(\gamma^\intercal, \beta^\intercal)$, the risk score for the future subject is

$$\mathbf{D}_0^\intercal \widehat{\gamma} + \sum_{l=1}^{r} \widehat{\beta}_l \mathbf{K}_{\mathbf{Z}_0}^\intercal \widehat{\boldsymbol{\zeta}}_l \widehat{\lambda}_l^{-\frac{1}{2}},$$

where $\mathbf{K}_{\mathbf{Z}_0} = (K(\mathbf{Z}_0, \mathbf{Z}_1), \ldots, K(\mathbf{Z}_0, \mathbf{Z}_n))^\intercal$.

2.4 General WLR and Kernel Selection

2.4.1 Testing and Prediction with General Weights

Although parameter estimates using the Gehan-weighted WLR are consistent, their performance in finite samples will vary depending on the underlying distributions of survival and censoring. To optimize the power of our pathway test and the accuracy of our risk prediction for a given dataset, it is desirable to extend these procedures to incorporate a general weight $\phi(\beta; \gamma, t)$. Since the WLR does not in general correspond to an objective function, we follow the same iterative strategy as proposed in Jin et al. (2003) to derive estimation and testing procedures for a general weight.

We first discuss estimation, using the computationally convenient kernel PCA approximation which uses a pre-specified proportion $\mathfrak{p}$ of the eigenvalues; results corresponding to the original kernel can be

obtained by setting $\mathfrak{p} = 1$. We initialize $\{\widehat{\gamma}^{(0)}, \widehat{\beta}^{(0)}\} = \{\widehat{\gamma}_G, \widehat{\beta}_G\}$, which are estimated using the Gehan weights as described in Section 2.3.3. Then at step k, from given estimates $\{\widehat{\gamma}^{(k-1)}, \widehat{\beta}^{(k-1)}\}$, we may obtain an updated estimator $\{\widehat{\gamma}^{(k)}, \widehat{\beta}^{(k)}\}$ as the minimizer of

$$\widetilde{L}_\phi^R(\beta;\gamma) = n^{-2}\sum_{i=1}^n\sum_{j=1}^n \psi(\widehat{\beta}^{(k-1)};\widehat{\gamma}^{(k-1)}, e_i(\widehat{\beta}^{(k-1)};\widehat{\gamma}^{(k-1)}))\Delta_i|\widetilde{e}_j(\beta;\gamma) - \widetilde{e}_i(\beta;\gamma)|_+ + \frac{c_k^2}{2}\beta^\intercal\beta \tag{2.14}$$

where the c_k are a sequence of regularizing constants which converge to some c, and $\psi(\beta;\gamma,t) = \phi(\beta;\gamma,t)/S^{(0)}(\beta;\gamma,t)$ for

$$S^{(k)}(\beta;\gamma,t) = n^{-1}\sum_{j=1}^n I\{e_j(\beta;\gamma) \geq t\}\begin{pmatrix}\mathbf{D}_j \\ \widetilde{\mathbf{B}}_{rj}\end{pmatrix}^{\otimes k}.$$

As $k \to \infty$, $\{\widehat{\gamma}^{(k)}, \widehat{\beta}^{(k)}\}$ converges to $\{\widehat{\gamma}_\phi, \widehat{\beta}_\phi\}$, a solution to $\widetilde{\mathbf{U}}_\phi(\beta;\gamma) + c^2\beta = 0$, where

$$\widetilde{\mathbf{U}}_\phi(\beta;\gamma) = n^{-1}\sum_{i=1}^n \phi(\beta;\gamma,e_i(\beta;\gamma))\Delta_i\left[\begin{pmatrix}\mathbf{D}_i \\ \widetilde{\mathbf{B}}_{ri}\end{pmatrix} - \frac{S^{(1)}(\beta;\gamma,e_i(\beta;\gamma))}{S^{(0)}(\beta;\gamma,e_i(\beta;\gamma))}\right].$$

For testing with a general weight, we may derive the score test using the same arguments as in Section 2.3 but add weights $\psi(\widehat{\beta}^{(k-1)};\widehat{\gamma}^{(k-1)}, e_i(\widehat{\beta}^{(k-1)};\widehat{\gamma}^{(k-1)}))$ to (2.9), where $\{\widehat{\gamma}^{(k-1)}, \widehat{\beta}^{(k-1)}\}$ are some initial estimates of $\{\gamma, \beta\}$. When testing $H_0 : h(\cdot) = 0$, it is convenient to set $\widehat{\beta}^{(k-1)} = 0$, its value under the null, and let $\widehat{\gamma}^{(k-1)} = \widetilde{\gamma}_\phi$, the WLR estimator under H_0 which can be found using the iterative method in Jin et al. (2003). With these choices, no iteration is needed to calculate the test statistic after $\widetilde{\gamma}_\phi$ has been found. Our WLR KM test statistic for a general weight ϕ is thus $\widehat{Q}_\phi = \widehat{\mathbf{U}}_\phi(\widetilde{\gamma}_\phi)^\intercal\widehat{\mathbf{U}}_\phi(\widetilde{\gamma}_\phi)$ where

$$\widehat{\mathbf{U}}_\phi(\gamma) = n^{-2}\sum_{i=1}^n\sum_{j=1}^n \psi(0;\gamma,e_i(0;\gamma))\Delta_i(\widetilde{\mathbf{B}}_{ni} - \widetilde{\mathbf{B}}_{nj})\mathbf{I}\{e_j(0;\gamma) \geq e_i(0;\gamma)\}.$$

Perturbations used to approximate the null take the form $\widehat{Q}_\phi^* = (\widehat{\mathbf{U}}_\phi^*(\widetilde{\gamma}_\phi^*) - \widehat{\mathbf{U}}_\phi(\widetilde{\gamma}_\phi))^\intercal(\widehat{\mathbf{U}}_\phi^*(\widetilde{\gamma}_\phi^*) - \widehat{\mathbf{U}}_\phi(\widetilde{\gamma}_\phi))$ where $\widetilde{\gamma}_\phi^*$ is a perturbation of $\widetilde{\gamma}_\phi$ under H_0 associated with a particular realization of $\mathcal{V}$, and where:

$$\widehat{\mathbf{U}}_\phi^*(\gamma) = n^{-2}\sum_{i=1}^n\sum_{j=1}^n \psi^*(0;\gamma,e_i(0;\gamma))\Delta_i(\widetilde{\mathbf{B}}_{ni} - \widetilde{\mathbf{B}}_{nj})\mathbf{I}\{e_j(0;\gamma) \geq e_i(0;\gamma)\}\mathcal{V}_i\mathcal{V}_j.$$

with $\psi^*(0;\gamma,t) = 1/S^{(0)*}(0;\gamma,t)$ for $S^{(0)*}(0;\gamma,t) = n^{-1}\sum_{j=1}^n \mathbf{I}\{e_j(0;\gamma) \geq t\}\mathcal{V}_j$. When $\phi = 1$, $\widehat{Q}_\phi$ reduces to the KM score statistic proposed in Cai et al. (2011) in the absence of clinical covariates.

2.4.2 Combining P-Values Across Models and Kernels

In practice, little prior information is available on the optimal choices of the kernel or WLR weights for a given dataset. Hence, data-driven methods to combine information from multiple weights and kernels are

desirable. To this end, we propose a test statistic $\widehat{A}$ which combines the test statistics for several weight-kernel combinations of interest. To generate the null distribution of $\widehat{A}$, we use a common collection of realizations $\{\mathcal{V}_{(1)}, \ldots, \mathcal{V}_{(B)}\}$ of the n-vector $\mathcal{V} = (\mathcal{V}_1, \ldots, \mathcal{V}_n)$ of iid mean-1 variance-1 random variables $\mathcal{V}_i$ to generate null distributions of each weight-kernel combination of interest, thus obtaining their joint null distribution. We can then generate the null distribution of $\widehat{A}$ by calculating its value, $\widehat{A}^*_{(b)}$, associated with each $\mathcal{V}_{(b)}$, and generate a P-value by comparing $\widehat{A}_{\text{obs}}$ to $\{\widehat{A}^*_{(1)}, \ldots, \widehat{A}^*_{(B)}\}$ – i.e., $\widehat{p}_{\text{combined}} = \#\{\widehat{A}^*_{(b)} \geq \widehat{A}_{\text{obs}}\}/B$.

We propose to use the method developed in Xu et al. (2003) for combining dependent tests. This method first transforms the P-values into z-scores by $Z_k = \Phi^{-1}(1 - p_k)$, yielding a vector of observed z-scores $\mathbf{Z}_{\text{obs}}$ and its associated null $\{\mathbf{Z}^*_{(1)}, \ldots, \mathbf{Z}^*_{(B)}\}$. The $\mathbf{Z}^*_{(b)}$ are used to estimate Γ, the covariance of $\mathbf{Z}$, and then the function $W(c) = \mathbf{Z}_c^\intercal \Gamma^{-1} \mathbf{Z}$ is computed for $c \in [0, 4]$, where $\mathbf{Z}_c$ is a vector with entries $\max\{Z_k, c\}$. By calculating $W(c)$ for both the observed $\mathbf{Z}_{\text{obs}}$ and the null $\mathbf{Z}^*_{(b)}$, an estimated P-value $p_{\text{Xu}}(c)$ can be calculated for each c, and the final test statistic whose value on the observed data is compared to its value on the null is $\widehat{A}_{\text{Xu}} = \min_{c \in [0,4]} p_{\text{Xu}}(c)$.

As in testing, the best choice of kernel and weight for risk prediction may not be obvious. We consider two automated approaches for determining which kernel-weight combination to use for the estimation of h. In the first approach, we apply the associated KM tests to the pathway and use the kernel-weight pair which yields the smallest p-value. In the second approach, we choose the kernel-weight pair which optimizes prediction accuracy. Specifically, for each pair, we estimate a C-statistic which captures how well the order of the associated risk predictions corresponds to the order of the true survival times T_i during a pre-specified follow-up period $(0, \tau)$, $C_\tau = P(\widehat{\gamma}^\intercal \mathbf{D}_i + \widehat{h}(\mathbf{Z}_i) > \widehat{\gamma}^\intercal \mathbf{D}_j + \widehat{h}(\mathbf{Z}_j) | T_i > T_j, T_j < \tau)$. We estimate C_τ in the presence of censoring using the nonparametric estimator proposed in Uno et al. (2011), and select the kernel-weight combination which yields the largest value of C_τ.

2.5 Simulation Studies

2.5.1 Testing

We conducted simulation studies to assess the performance of the proposed testing and estimation procedures. We generated the pathway covariates $\mathbf{Z}$ from a multivariate normal distribution with mean 0 and compound symmetry covariance structure with variance 1 and correlation $\wp$. We considered pathways of

size $P = 5$ and 20, and correlations $\wp = 0.8, 0.5$ and 0.2 to represent strong, moderate, and weak within-pathway correlation. For simplicity we did not include any additional covariates. We simulated different types of underlying signals $h(\mathbf{z})$ to understand the performance of our procedures in different settings. For a given $h(\mathbf{z})$, we generated survival times according to the model $\log T = h(\mathbf{Z}) + E$. We compared two different types of error distributions, generating E from either the extreme value distribution (EVD) or a logistic distribution Logistic(μ, σ), where μ and σ were chosen so that the survival times resulting from both choices of E were of comparable magnitude. The censoring was generated from a uniform distribution with range chosen so that approximately 25% of the individuals were censored.

To assess the performance of the testing procedures in each setting, we simulated 2000 data sets for empirical size calculations and 1000 for empirical power, and compared the AFT KM pseudo-score tests with Gehan weights (WLR_{Ge}) and log-rank weights (WLR_{LR}) for the Gaussian, quadratic, and linear kernels. We note that the WLR_{LR} test statistic is the same as the Cox KM test statistic proposed by Cai et al. (2011) when there are no other covariates, although the specific method used for approximating the null distribution differs. All null distributions were generated using $B = 2000$ perturbations, and kernel PCA was used to reduce dimensionality, with $\mathfrak{p} = 0.95$. For the linear kernel which needs no tuning, we considered the perturbation p-value $\widehat{p}$ and the χ^2-approximated p-value $\widehat{p}_{\chi^2}$ directly. For the kernels which rely on a tuning parameter ρ, we considered the test based on $\widehat{T}_{\mathcal{I}} = \inf_{\rho \in \mathcal{I}} \left\{ \widehat{p}_{\chi^2}(\rho) \right\}$ and compared two other candidate test statistics: $\widehat{S}_{\mathcal{I}} = \sup_{\rho \in \mathcal{I}} \left\{ \widehat{Q}(\rho)/\widehat{\sigma}(\rho) \right\}$ and $\widehat{R}_{\mathcal{I}} = \inf_{\rho \in \mathcal{I}} \left\{ \widehat{p}(\rho) \right\}$, where $\widehat{Q}(\rho)$, $\widehat{p}(\rho)$, and $\widehat{p}_{\chi^2}(\rho)$ denote respectively the test statistic, p-value from perturbation, and p-value from the χ^2 approximation derived under kernel function $K(\cdot, \cdot; \rho)$; and $\widehat{\sigma}(\rho)$ is the estimated standard error of $\widehat{Q}(\rho)$ obtained from the perturbations. The results for the different tests used for each kernel were nearly identical, so those based on the χ^2-approximations are shown in the tables. We also considered the omnibus test described in Section 2.4.2; for comparison, we also considered three other candidates $\widehat{A}$ upon which to base an omnibus test: $\widehat{A}_{\text{Fisher}} = -2 \sum_{k=1}^{K} \log p_k$ (Fisher, 1925); $\widehat{A}_{\text{Truncated}} = -2 \sum_{k=1}^{K} \mathbf{I}\{p_k \leq 0.05\} \log p_k$ (Zaykin et al., 2002); and $\widehat{A}_{\text{min-p}} = \min_k \{p_k\}$. We considered combining across kernels within WLR weights, across weights within kernel, and across all six weight-kernel combinations; for simplicity, only the p-values combined across all six are shown in the tables.

For comparison to our KM-based tests, we also considered a method for assigning a p-value to a pathway based on the marginal associations between the genes in the pathway and survival. For each gene Z_i, $i = 1, \ldots, P$, in the pathway, we calculated a marginal p-value $\widehat{p}_i^{\text{marg}}$ from the Wald test from a standard

univariate Cox model, and then calculated the minimum of these P p-values, $\widehat{p}^{\text{marg}}_{\min} = \min\{\widehat{p}^{\text{marg}}_1, \ldots, \widehat{p}^{\text{marg}}_P\}$. We adjusted this p-value for multiple testing using the effective number of tests, M_{eff}, as described in Nyholt (2004). That is, we define $\widehat{p}^{\text{marg}}_{\text{pathway}} = 1 - [1 - \widehat{p}^{\text{marg}}_{\min}]^{M_{\text{eff}}}$, where for observed eigenvalues $\widehat{\ell}_j$ from the covariance matrix of the genes in the pathway, $M_{\text{eff}} = 1 + (P-1)(1 - \text{Var}(\widehat{\ell}_1, \ldots, \widehat{\ell}_P)/P)$.

To examine the validity of the test procedure in finite samples, we generated data under the null setting $h(\mathbf{z}) = 0$. The empirical sizes at Type I error rate of 0.05 are shown in Table 2.1 for $n = 200$ and 400. When $n = 200$, the empirical sizes of the KM tests tend to be slightly below their nominal level, especially when the within-pathway correlation is low. The four combined p-values maintain their nominal level. The empirical sizes of all the tests are closer to their nominal level when the sample size is increased to $n = 400$.

To assess the power of the proposed tests, we considered for $n = 200$ (1) a linear signal, $h_1(\mathbf{z}) = c(z_1 + z_2 + z_3 + z_4 + z_5)$ with $c = 0.05$, and (2) a nonlinear signal, $h_2(\mathbf{z}) = c[z_1 + 4z_1^2 + z_2 + 4z_2^2 - 2z_1z_2 + g(z_3)(4z_4 + 4z_5) + (1 - g(z_3))(-3z_4 - 3z_5 + 4z_4z_5)]$, with $c = 0.75$ and $g(z) \sim \text{Bernoulli}(e^{-|z|})$. This function was chosen to have different types of nonlinear signal: the linear, quadratic, and interactive effects of z_1 and z_2, and the latent classes defined by z_3 with differential signal defined by z_4 and z_5. Results are shown in Table 2.2.

When the true signal is linear, one would expect the linear kernel to outperform the other kernels, but interestingly, all three kernels yield tests with competitive power. This can in part be attributed to the fact that both the Gaussian and quadratic kernels can capture primarily linear effects at certain values of their tuning parameters. When the true signal is nonlinear, the linear kernel performs poorly, particularly in the high correlation setting. In comparison, the quadratic kernel has somewhat higher power, while the Gaussian kernel can have substantially more power. For all tests, the power decreases somewhat when we increase the number of covariates from $P = 5$ to $P = 20$; however, the power loss is small when the correlation is high due to the low effective degrees of freedom in such settings. This highlights one of the advantages of the KM based tests – that they benefit from within-pathway correlation. We do not see substantial gains from using the quadratic kernel, either in these setting or in settings where the quadratic effect is stronger (results not shown).

Among the four omnibus tests considered, the Xu method appears to most robustly maintain power across different settings. Its performance is slightly weaker than the other omnibus tests when the true signal is linear, but it still outperforms the worst choices of weight-kernel combination; however, in the nonlinear setting, it tends to be more powerful than the other tests. In several of the nonlinear settings, its

Table 2.1: Empirical sizes (%) at Type I error rate of 0.05 when $n = 200$ and $n = 400$. Testing was performed using the Gaussian, linear, and quadratic kernels using both the WLR_{Ge} and WLR_{LR} tests. For the linear kernel, the χ^2 approximated p-value is presented; for the Gaussian and quadratic kernels, the p-value based on the χ^2 $\min p$ statistic ($\widehat{T}_{\mathcal{I}}$ in the text) is presented. Results shown use kernel PCA with 95% of eigenvalues used. Also shown are the empirical sizes of the four proposed omnibus tests – Fisher, Truncated, min-p, and Xu – across all 6 weight-kernel combinations. Shown for comparison are results for the marginal gene method.

Correlation		0.2				0.5				0.8			
Pathway Size		5		20		5		20		5		20	
Error		EVD	logis	EVD	logis	EVD	logis	EVD	logis	EVD	logis	EVD	logis
$n = 200$													
Marginal Gene		5	5	5	5	5	5	5	5	7	6	6	6
WLR_{Ge}	Gaussian	3	3	3	3	4	3	4	3	3	4	4	4
	Linear	2	2	2	2	4	3	3	3	3	3	4	4
	Quadratic	3	3	3	3	4	4	4	3	3	3	3	4
WLR_{LR}	Gaussian	3	4	4	3	4	4	5	5	4	4	5	4
	Linear	3	3	2	2	4	4	4	4	4	4	4	4
	Quadratic	4	4	4	3	4	4	5	5	4	4	5	4
Omnibus	Fisher	3	3	3	3	4	4	4	4	3	4	4	4
	Truncated	3	3	3	2	4	4	4	3	3	4	4	4
	Min-P	3	3	2	2	4	3	4	3	2	3	3	3
	Xu	2	2	2	2	4	3	4	3	2	3	3	4
$n = 400$													
Marginal Gene		5	5	5	5	6	6	6	5	6	6	5	5
WLR_{Ge}	Gaussian	4	4	4	4	5	4	6	4	5	4	4	4
	Linear	3	4	3	3	5	4	5	4	4	4	4	4
	Quadratic	4	4	4	4	5	5	6	5	4	4	4	4
WLR_{LR}	Gaussian	4	5	4	5	5	5	6	5	5	5	5	4
	Linear	4	4	3	4	5	4	6	4	4	5	5	4
	Quadratic	4	5	4	5	5	5	6	5	5	6	6	4
Omnibus	Fisher	3	4	4	5	5	5	6	5	4	4	5	4
	Truncated	3	4	4	4	5	4	6	5	4	4	5	4
	Min-P	3	4	4	3	5	4	6	4	4	4	5	4
	Xu	3	3	3	3	5	4	6	4	4	4	5	4

Table 2.2: Empirical power (%) for linear and nonlinear signal at Type I error rate of 0.05 when $n = 200$. Tests shown are identical to those in Table 2.1.

Correlation		0.2				0.5				0.8			
Pathway Size		5		20		5		20		5		20	
Error		EVD	logis	EVD	logis	EVD	logis	EVD	logis	EVD	logis	EVD	logis
linear $h(\mathbf{z})$													
Marginal Gene		27	20	18	15	57	43	47	35	77	66	74	58
WLR_{Ge}	Gaussian	24	25	17	23	46	48	42	43	55	62	60	60
	Linear	22	23	12	17	46	48	39	39	62	66	62	62
	Quadratic	27	26	17	23	46	49	42	43	52	56	61	60
WLR_{LR}	Gaussian	35	22	25	20	57	42	55	40	70	57	72	56
	Linear	30	19	20	14	58	43	52	38	74	62	74	59
	Quadratic	36	23	26	20	59	44	55	40	66	52	72	56
Omnibus	Fisher	31	25	21	20	56	48	50	42	70	66	70	63
	Truncated	30	23	20	20	56	48	50	42	70	64	71	62
	Min-P	30	23	19	18	52	45	49	40	63	58	68	58
	Xu	25	20	16	15	51	42	48	38	64	59	66	57
nonlinear $h(\mathbf{z})$													
Marginal Gene		52	54	37	40	44	48	34	37	18	19	10	10
WLR_{Ge}	Gaussian	87	88	28	30	85	87	23	25	93	92	51	50
	Linear	38	38	23	26	19	19	13	16	5	5	4	4
	Quadratic	39	40	27	29	32	34	14	17	67	70	37	36
WLR_{LR}	Gaussian	94	95	41	44	98	99	48	46	100	100	94	94
	Linear	47	49	34	38	31	32	25	26	4	5	4	5
	Quadratic	50	51	38	41	52	55	29	30	95	95	73	73
Omnibus	Fisher	67	69	32	36	62	65	25	27	90	90	44	44
	Truncated	71	74	33	37	73	72	27	28	97	95	56	57
	Min-P	92	92	33	36	96	97	40	39	100	99	86	85
	Xu	98	98	33	36	100	99	66	68	100	100	98	97

power is greater than all individual weight-kernel tests, suggesting it is gaining power when several tests have nearly significant results. This points to the robustness of the omnibus test, which could make it useful in a wide range of settings.

2.5.2 Estimation

To assess the performance of the estimation procedure when the true signal is linear (h_1 from above, with $c = 0.4$) or nonlinear (h_2 with $c = 3$), we ran 500 simulations in the same configurations of correlations, pathway sizes, and error distributions. Each simulation consisted of two data sets: a training data set with sample size $n_{\text{train}} = 100$ on which to build the estimates $\widehat{h}$, and a validation data set with sample size $n_{\text{test}} = 1000$ on which to assess how well each estimate predicted survival up until some time t_0 using the C-statistic (Uno et al., 2011). In these simulations, we selected a reference time t_0 that was near the 70^{th} percentile of followup time.

Using the training data, we built estimators for the Gehan and LR weights using linear, quadratic, and Gaussian kernels. We also fit standard full Cox and AFT models for comparison. We applied these estimators and models to the validation data and calculated the C-statistic. We considered two methods for choosing which kernel to use in the final estimate $\widehat{h}$; in one, we choose the kernel with the smallest p-value from the pseudo-score test in the training data; in the other, we choose the kernel yielding the largest estimated C-statistic in the training data. Results are presented in Table 2.3.

In the linear setting, the C-statistics are nearly identical across all models. In the nonlinear setting, we see huge gains over the full model by using nonlinear kernel functions. In this setting, the predictive ability of the three kernels varies more, so it is more meaningful to select a kernel. The model using the kernel selected using the C-statistic has almost uniformly better predictive ability than all other models, and thus, we recommend selecting kernel based on the estimated C-statistic.

2.6 Example: Breast Cancer Gene Expression Study

Genomic information has already improved our understanding of breast cancer. The mutations found in the BRCA1 and BRCA2 genes identify women at high risk of developing breast cancer (Narod and Foulkes, 2004), and a number of gene expression signatures have been introduced into clinical practice to better identify cancers with high and low risk of recurrence (Desmedt et al., 2011). Despite these advances,

Table 2.3: Empirical C-statistic (%) for predicting survival up to time t_0 for linear and nonlinear signal, where t_0 is approximately the 70$^{\text{th}}$ percentile of follow-up time. $\widehat{h}(\mathbf{z})$ is built in a training data set with $n_{\text{train}} = 100$; results shown use kernel PCA with 95% of eigenvalues used. All C-statistics are calculated by applying $\widehat{h}(\mathbf{z})$ to a testing data set with $n_{\text{test}} = 1000$. For kernel methods, we present the C-statistic for $\widehat{h}$ estimated from each of the three kernels, and the C-statistic if we select kernel based on KM test p-value P, or $\widehat{C}$, the estimated C-statistic in the training data. For comparison, we present the Cox and AFT full models, which are the models fit with all pathway variables as linear covariates.

Correlation		0.2				0.5				0.8			
Pathway Size		5		20		5		20		5		20	
Error		EVD	Exp	EVD	Exp	EVD	Exp	EVD	Exp	EVD	Exp	EVD	Exp
linear $h(\mathbf{z})$													
Full	AFT-G	76	75	73	74	79	79	76	77	82	83	79	78
	AFT-LR	77	75	74	73	79	79	77	77	82	83	80	79
	Cox	77	75	74	73	79	79	77	77	82	82	80	79
WLR_{Ge} Kernel	Gaussian	77	75	73	73	79	79	78	78	82	82	82	81
	Linear	77	75	74	74	80	79	78	79	82	83	82	81
	Quadratic	77	75	73	73	79	79	77	78	82	83	82	81
WLR_{LR} Kernel	Gaussian	77	75	73	73	79	79	78	78	82	82	82	81
	Linear	77	75	75	74	80	79	78	79	82	83	83	81
	Quadratic	77	75	73	73	80	79	78	78	82	83	83	81
Omnibus P		77	75	75	74	79	79	78	79	82	83	82	81
Omnibus C		77	75	74	74	79	79	78	79	82	83	82	81
nonlinear $h(\mathbf{z})$													
Full	AFT-G	55	55	52	53	54	53	53	53	55	55	53	53
	AFT-LR	55	55	52	53	54	54	53	53	55	55	53	53
	Cox	55	55	52	53	54	54	53	53	55	55	53	53
WLR_{Ge} Kernel	Gaussian	66	68	52	53	65	65	53	54	67	67	60	61
	Linear	55	55	52	53	54	53	52	52	53	52	51	51
	Quadratic	56	56	52	53	59	59	53	54	62	61	57	58
WLR_{LR} Kernel	Gaussian	66	69	53	54	66	66	55	56	67	69	63	63
	Linear	55	55	52	53	54	54	52	53	53	53	51	51
	Quadratic	56	57	52	53	60	60	54	55	62	63	59	60
Omnibus P		66	68	53	54	66	65	55	55	66	68	61	62
Omnibus C		67	69	53	54	67	67	55	56	68	68	62	62

approximately 60% of patients with early-stage breast cancer are given adjuvant therapy in addition to local treatment, while only a small proportion are thought to benefit (Reis-Filho and Pusztai, 2011). Better markers of aggressive disease would help physicians predict which patients could safely avoid adjuvant therapy and its negative side-effects, and which patients should be treated with more aggressive therapy.

One promising hypothesis-driven approach is to investigate the effects of candidate pathways on patient survival. We examine the associations between recurrence-free survival and 32 candidate pathways from the molecular signature database. We consider, for example, the p53 pathway because mutation of p53 has been previously found to be associated with more aggressive disease and worse overall survival in breast cancer (Gasco et al., 2002).

To assess the effects of these pathways on breast cancer progression, we applied our pseudo-score test to each of the pathways in a training set of 286 lymph node negative breast cancer patients who received no systemic adjuvant therapy (Wang et al., 2005). We used tumor gene expression assessed on the Affymetrix U133a Gene Chip, and controlled for ER status. A total of 107 deaths or recurrences were observed, with follow-up time ranging between 2 months and 14.3 years (median 7.2 years); 63% of observations were censored. Figure 2.1 shows the results of the testing procedure. For each pathway, we compare the p-value from the marginal-gene based method to the p-value for the Xu omnibus test combining KM tests for all six weight-kernel combinations. 20 pathways were significant at the nominal 0.05 level using the marginal gene method, and 22 were significant using the omnibus test.

Since multiple pathways are under consideration, we need to adjust for multiple comparisons. To adjust the marginal gene-based p-value we may use the same $\widehat{p}^{\text{marg}}_{\text{pathway}}$ defined in Section 2.5.1, but with the effective degrees of freedom M_{eff} calculated using all the genes in the pathways under consideration. To adjust the KM-based p-value, we may use the approximate null distribution generated by perturbation to get the null distribution of the minimal p-value across the 32 pathway tests under the null, and compare the observed pathway p-values to this null distribution. After adjustment, one pathway is still significant using the marginal gene method, while 5 remain significant using the omnibus KM test.

For each of the five pathways declared most significant by our KM test, we estimated the pathway effect on recurrence-free survival in the training set using the kernel which optimized the C-statistic; for comparison, we also fit a standard Cox model. We applied these estimates to an independent validation set of 119 lymph node negative patients with no adjuvant therapy, with gene expression assessed on the same chip (Sotiriou et al., 2006). In this validation set, 27 deaths or recurrences were observed, with follow-

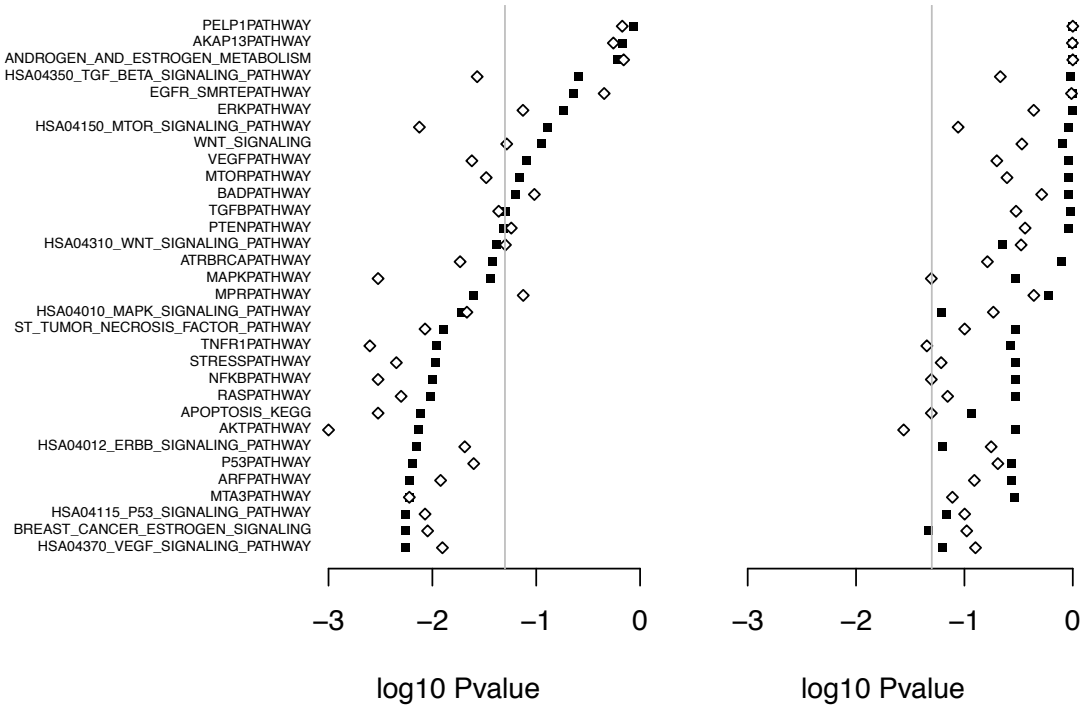

Figure 2.1: $\log_{10}$ p-values for testing the overall effect of 32 pathways on breast cancer survival. Black squares represent the pathway p-values from the marginal-gene approach, and the pathways are ordered by this p-value. Diamonds represent the Xu combined KM p-values. The left hand panel shows the unadjusted p-values, while the right hand panel shows the p-values adjusted for multiple testing. Results are based on $B = 5000$ perturbations.

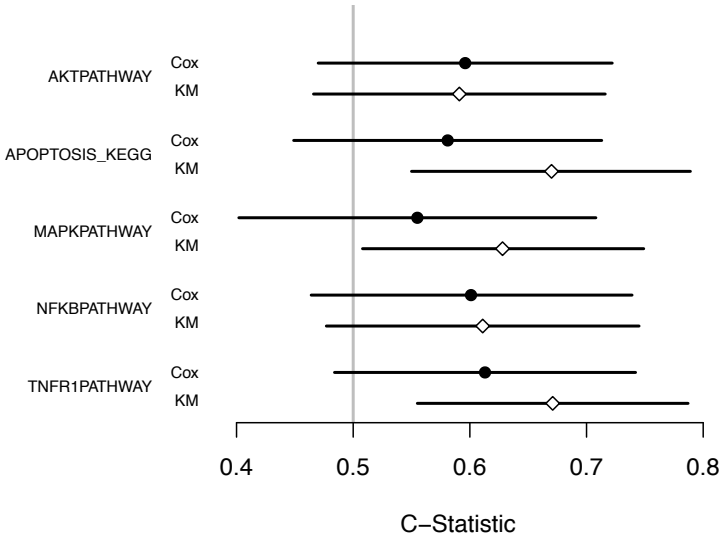

Figure 2.2: Point estimates and confidence intervals for the C-statistics for a full standard Cox model (Cox) and our KM model (KM) for the 5 pathways with lowest KM p-value. Models were built on training data ($n = 286$) and applied to testing data ($n = 119$), where the C-statistics and confidence intervals were estimated.

up time ranging between 2 months and 14.5 years (median 7.7 years); 77% of observations were censored. We assessed the accuracy of each model using the C-statistic and calculated 95% confidence intervals (Uno et al., 2011). The confidence intervals for the C-statistics for the Cox model fit to each pathway crossed 50%, while three of the pathways had KM confidence intervals which did not cross 50%, suggesting potential for improvement in risk prediction using AFT KM modeling.

2.7 Discussion

In this paper, we proposed KM based procedures to test for the effect of pathways on survival in an AFT model framework. By taking a pathway based approach, we take advantage of pre-existing biological knowledge and privilege groups of genes believed to work together. By working with both linear and nonlinear kernels, we can detect linear effects well, while improving power to detect nonlinear effects when compared to methods based on linearity assumptions. When interest lies in risk prediction, our proposed risk scores from the AFT KM framework may be easily used for risk assessment. These risk scores have the potential to capture nonlinear effects which can improve over methods assuming linearity, as demonstrated in simulation. They also gain strength from within-pathway correlation, which is likely to exist in gene expression data.

In analysis of real data, it is unlikely that the researcher knows which weight or kernel is most appropriate for his or her data set. Thus, we have proposed omnibus testing and estimation procedures that allow the data to drive weight and kernel choice. The perturbation procedures used to generate the null distributions for our AFT KM tests enable us to efficiently combine information across kernels and weights. Based on simulation results for testing, the power lost is often minimal, but the power gained over the worst choice can be quite large. Moreover, for risk prediction, it appears that the omnibus estimation procedure has negligible loss in prediction accuracy when compared to the optimal kernel. Our perturbation procedures also allow for straightforward control of multiple testing that is not overly stringent when there is between-pathway correlation.

When the underlying signal is sparse with only a few genes in a pathway associated with survival, it would also be interesting to extend our proposed procedure to allow for feature selection under the KM framework. We were motivated to develop these kernel machine methods for the AFT model because of an attractive property of linear models that when two groups of covariates are independent, their marginal

effects on outcome are equal to their joint effects. In the context of gene-sets, this would mean that if we are considering high-dimensional genetic data which we can divide into a large number of independent pathways, we may assess the effect of each pathway individually, and then combine their marginal effects additively into a joint model. Future work will explore this application to large data sets, and investigate our abilities to combine information across both independent and correlated pathways.

2.8 Appendix: Asymptotic Distribution of the Test Statistic

Here we derive the asymptotic null distribution of our test statistic. Throughout, we assume that the covariates $\mathbf{D}_i$ and the genomic marker values $\mathbf{Z}_i$ are bounded by a constant $\mathbf{z}_m$. We assume that the true value of γ, γ_0, is an interior point of a compact set Ω, and without loss of generality, we also assume that X has a finite support $[0, \tau]$. For ease of notation, unless otherwise noted, the supremum is always taken over $(-\infty, \log(\tau)]$ for the index t, $[-\mathbf{z}_m, \mathbf{z}_m]$ for $\mathbf{Z}$ and Ω for γ. We also require the same set of assumptions given in Jin et al. (2003) for the iterative WLR estimation procedures. For simplicity, we focus on the Gehan weight but note that similar arguments can be used for the general weight. From Jin et al. (2001, 2003), we have $n^{\frac{1}{2}}(\widetilde{\gamma} - \gamma_0) = n^{-\frac{1}{2}} \sum_{i=1}^{n} \mathcal{U}_{\gamma i} + o_p(1)$ for some independent and identically distributed random variables $\mathcal{U}_{\gamma i}$, where $\widetilde{\gamma} = \widetilde{\gamma}_G$ in the text. We will derive the null distribution of $\widehat{Q}(\widetilde{\gamma}; \rho)$ and demonstrate its convergence as a process in ρ. The kernel function K is assumed to be continuously differentiable and the regularity conditions required in Braun (2005) are assumed to hold for the convergence of the empirical eigenvalues and eigenfunctions of $\mathcal{H}_K$. Under these regularity conditions, the convergence of the null distribution of $\widetilde{Q}(\widetilde{\gamma}; \rho)$, the test statistic with kernel PCA, can be derived using the convergence of $\widetilde{\mathbb{K}}$ to the kernel matrix corresponding to a truncated kernel, which spans $\mathcal{H}_{K_{r_0}} = \text{span}\{\sqrt{\lambda_1}\zeta_1(\cdot), \ldots, \sqrt{\lambda_{r_0}}\zeta_{r_0}(\cdot)\}$(Braun, 2005).

The test statistic takes the form $\widehat{Q}(\widetilde{\gamma}, \rho) = \widehat{\mathbf{R}}(\widetilde{\gamma})^{\intercal} \mathbb{K}(\rho) \widehat{\mathbf{R}}(\widetilde{\gamma})$, where $\widehat{\mathbf{R}}(\gamma) = (\widehat{R}_1(\gamma), \ldots, \widehat{R}_n(\gamma))^{\intercal}$,

$$\widehat{R}_i(\gamma) = n^{-2} \sum_{j=1}^{n} \left\{ \int Y_j(t; \gamma) dN_i(t; \gamma) - \int Y_i(t; \gamma) dN_j(t; \gamma) \right\},$$

$N_i(t; \gamma) = \Delta_i \mathbf{I}[e_i(0; \gamma) \leq t]$ and $Y_i(t; \gamma) = I[e_i(0; \gamma) \geq t]$. We can write $n\widehat{Q}(\widetilde{\gamma}, \rho)$ as:

$$\begin{aligned} n\widehat{Q}(\widetilde{\gamma}, \rho) &= n \sum_{i=1}^{n} \sum_{j=1}^{n} K(\mathbf{Z}_i, \mathbf{Z}_j; \rho) \widehat{R}_i(\widetilde{\gamma}) \widehat{R}_j(\widetilde{\gamma}) \\ &= \int \int K(\mathbf{u}, \mathbf{v}; \rho) d\left\{ \sqrt{n}\, \widehat{\mathbb{W}}_R(\mathbf{u}; \widetilde{\gamma}) \right\} d\left\{ \sqrt{n}\, \widehat{\mathbb{W}}_R(\mathbf{v}; \widetilde{\gamma}) \right\}. \end{aligned}$$

$$
\begin{aligned}
\text{where} \qquad \widehat{\mathbb{W}}_R(\mathbf{u};\widetilde{\gamma}) &= \sum_{i=1}^{n} \mathbf{I}[\mathbf{Z}_i \le \mathbf{u}]\widehat{R}_i(\widetilde{\gamma}). \\
&= n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n} \mathbf{I}[\mathbf{Z}_i \le \mathbf{u}]\left\{\int Y_j(t;\widetilde{\gamma})dN_i(t;\widetilde{\gamma}) - \int Y_i(t;\widetilde{\gamma})dN_j(t;\widetilde{\gamma})\right\} \\
&= \int \widehat{\pi}(\mathbf{z}_m,t;\widetilde{\gamma})\widehat{\mathbf{W}}_N(\mathbf{u},dt;\widetilde{\gamma}) - \int \widehat{\pi}(\mathbf{u},t;\widetilde{\gamma})\widehat{\mathbf{W}}_N(\mathbf{z}_m,dt;\widetilde{\gamma}), \\
\widehat{\pi}(\mathbf{u},t;\gamma) &= n^{-1}\sum_{i=1}^{n} \mathbf{I}[\mathbf{Z}_i \le \mathbf{u}]Y_i(t;\gamma), \qquad \widehat{\mathbf{W}}_N(\mathbf{u},t;\gamma) = n^{-1}\sum_{i=1}^{n} \mathbf{I}[\mathbf{Z}_i \le \mathbf{u}]N_i(t;\gamma),
\end{aligned}
$$

and $\mathbf{I}(\mathbf{Z}_i \le \mathbf{u}) = I(Z_{i1} \le u_1, ..., Z_{ip} \le u_p)$. Let $\pi(\mathbf{u},t;\gamma) = E\left[\mathbf{I}[\mathbf{Z} \le \mathbf{u}]Y(t;\gamma)\right]$, $\mu(\mathbf{u},t;\gamma) = E\left[\mathbf{I}[\mathbf{Z} \le \mathbf{u}]N(t,\gamma)\right]$, $\widehat{\mathbb{W}}_N(\mathbf{u},t;\gamma) = \widehat{\mathbf{W}}_N(\mathbf{u},t;\gamma) - \mu(\mathbf{u},t;\gamma)$, and $\widehat{\mathbb{W}}_\pi(\mathbf{u},t;\gamma) = \widehat{\pi}(\mathbf{u},t;\gamma) - \pi(\mathbf{u},t;\gamma)$. Then we can expand $\sqrt{n}\,\widehat{\mathbb{W}}_R(\mathbf{u};\widetilde{\gamma})$ as:

$$
\int \widehat{\mathbb{W}}_\pi(\mathbf{z}_m,t;\widetilde{\gamma})\left\{\sqrt{n}\,\widehat{\mathbb{W}}_N(\mathbf{u},dt;\widetilde{\gamma})\right\} - \int \widehat{\mathbb{W}}_\pi(\mathbf{u},t;\widetilde{\gamma})\left\{\sqrt{n}\,\widehat{\mathbb{W}}_N(\mathbf{z}_m,dt;\widetilde{\gamma})\right\} \tag{2.15}
$$

$$
+\int \pi(\mathbf{z}_m,t;\widetilde{\gamma})\left\{\sqrt{n}\,\widehat{\mathbb{W}}_N(\mathbf{u},dt;\widetilde{\gamma})\right\} - \int \pi(\mathbf{u},t;\widetilde{\gamma})\left\{\sqrt{n}\,\widehat{\mathbb{W}}_N(\mathbf{z}_m,dt;\widetilde{\gamma})\right\} \tag{2.16}
$$

$$
+\int \left\{\sqrt{n}\,\widehat{\mathbb{W}}_\pi(\mathbf{z}_m,t;\widetilde{\gamma})\right\}\mu(\mathbf{u},dt;\widetilde{\gamma}) - \int \left\{\sqrt{n}\,\widehat{\mathbb{W}}_\pi(\mathbf{u},t;\widetilde{\gamma})\right\}\mu(\mathbf{z}_m,dt;\widetilde{\gamma}) \tag{2.17}
$$

$$
+\sqrt{n}\left\{\int \pi(\mathbf{z}_m,t;\widetilde{\gamma})\mu(\mathbf{u},dt;\widetilde{\gamma}) - \int \pi(\mathbf{u},t;\widetilde{\gamma})\mu(\mathbf{z}_m,dt;\widetilde{\gamma})\right\} \tag{2.18}
$$

We first show that the first pair of integrals (2.15) is $o_p(1)$. To this end, we note that by a functional central limit theorem (FCLT) (Pollard, 1990), $\sqrt{n}\,\widehat{\mathbb{W}}_N(\mathbf{u},t,\gamma)$ converges weakly to a Gaussian process in $(\mathbf{u},t,\gamma)$, denoted by $\mathbb{W}_N(\mathbf{u},t;\gamma)$. It follows that

$$
\sqrt{n}\,\widehat{\mathbb{W}}_N(\mathbf{u},t;\widetilde{\gamma}) = \sqrt{n}\,\widehat{\mathbb{W}}_N(\mathbf{u},t,\gamma_0) + o_p(1) \tag{2.19}
$$

by stochastic equicontinuity. On the other hand, by a uniform law of large numbers (ULLN) (Pollard, 1990), $\sup_{\mathbf{u},t,\gamma}|\widehat{\mathbb{W}}_\pi(\mathbf{u},t;\gamma)| = o_p(1)$ which implies that $\sup_{\mathbf{u},t}|\widehat{\mathbb{W}}_\pi(\mathbf{u},t;\widetilde{\gamma})| = o_p(1)$. This, together with Lemma A.3 of Bilias et al. (1997) and the strong representation theorem, implies that (2.15) $= o_p(1)$ uniformly in $\mathbf{u}$.

The integrals in (2.16) have the same limiting distribution as

$$
\int \pi(\mathbf{z}_m,t;\gamma_0)\left\{\sqrt{n}\,\widehat{\mathbb{W}}_N(\mathbf{u},dt;\gamma_0)\right\} - \int \pi(\mathbf{u},t;\gamma_0)\left\{\sqrt{n}\,\widehat{\mathbb{W}}_N(\mathbf{z}_m,dt;\gamma_0)\right\}.
$$

We may see this by adding and subtracting $\pi(\mathbf{u},t;\gamma_0)$ to and from the integrands $\pi(\mathbf{u},t;\widetilde{\gamma})$, and using the fact that $\pi(\mathbf{u},t;\widetilde{\gamma}) - \pi(\mathbf{u},t;\gamma_0) \xrightarrow{P} 0$, and by using (2.19) to replace the integrating functions by $\sqrt{n}\,\widehat{\mathbb{W}}_N(\mathbf{u},dt;\gamma_0)$.

The integrals in (2.17) have the same limiting distribution as

$$
\int \left\{\sqrt{n}\,\widehat{\mathbb{W}}_\pi(\mathbf{z}_m,t;\gamma_0)\right\}\mu(\mathbf{u},dt;\gamma_0) - \int \left\{\sqrt{n}\,\widehat{\mathbb{W}}_\pi(\mathbf{u},t;\gamma_0)\right\}\mu(\mathbf{z}_m,dt;\gamma_0)
$$

To see this, note that a FCLT implies that $\sqrt{n}\,\widehat{\mathbb{W}}_\pi(\mathbf{u},t,\gamma)$ converges weakly to a zero-mean Gaussian process $\mathbb{W}_\pi(\mathbf{u},t,\gamma)$. Hence, $\sqrt{n}\widehat{\mathbb{W}}_\pi(\mathbf{u},t;\widetilde{\gamma}) = \sqrt{n}\widehat{\mathbb{W}}_\pi(\mathbf{u},t;\gamma_0) + o_p(1)$, which allows us to replace the integrands $\sqrt{n}\widehat{\mathbb{W}}_\pi(\mathbf{u},t;\widetilde{\gamma})$ by $\sqrt{n}\widehat{\mathbb{W}}_\pi(\mathbf{u},t;\gamma_0)$. Further, we can expand

$$\begin{aligned}\int \left\{\sqrt{n}\,\widehat{\mathbb{W}}_\pi(\mathbf{u}_1,t;\gamma_0)\right\}\mu(\mathbf{u}_2,dt;\widetilde{\gamma}) \;&=\; \int \left\{\sqrt{n}\,\widehat{\mathbb{W}}_\pi(\mathbf{u}_1,t;\gamma_0)\right\}\mu(\mathbf{u}_2,dt;\gamma_0) &(2.20)\\ &\quad+\int \left\{\sqrt{n}\,\widehat{\mathbb{W}}_\pi(\mathbf{u}_1,t;\gamma_0)\right\}\{\mu(\mathbf{u}_2,dt;\widetilde{\gamma})-\mu(\mathbf{u}_2,dt;\gamma_0)\} &(2.21)\end{aligned}$$

and see that $(2.21) = o_p(1)$ by another application of Lemma A.3 of Bilias et al. (1997) and the strong representation theorem, because $\mu(\mathbf{u},dt;\widetilde{\gamma})-\mu(\mathbf{u},dt;\gamma_0)\xrightarrow{P}0$.

Finally, we can write (2.18) as:

$$\begin{aligned}\sqrt{n}\Big\{ &\int \pi(\mathbf{z}_m,t;\gamma_0)\mu(\mathbf{u},dt;\gamma_0) - \int \pi(\mathbf{u},t;\gamma_0)\mu(\mathbf{z}_m,dt;\gamma_0) &(2.22)\\ &+\int \pi(\mathbf{z}_m,t;\widetilde{\gamma})\{\mu(\mathbf{u},dt;\widetilde{\gamma})-\mu(\mathbf{u},dt;\gamma_0)\} &(2.23)\\ &+\int \{\pi(\mathbf{z}_m,t;\widetilde{\gamma})-\pi(\mathbf{z}_m,t;\gamma_0)\}\mu(\mathbf{u},dt;\gamma_0) &(2.24)\\ &-\int \pi(\mathbf{u},t;\widetilde{\gamma})\{\mu(\mathbf{z}_m,dt;\widetilde{\gamma})-\mu(\mathbf{z}_m,dt;\gamma_0)\} &(2.25)\\ &-\int \{\pi(\mathbf{u},t;\widetilde{\gamma})-\pi(\mathbf{u},t;\gamma_0)\}\mu(\mathbf{z}_m,dt;\gamma_0)\quad\Big\} &(2.26)\end{aligned}$$

The first line (2.22) is identically 0 because $\int \mathbf{I}[\mathbf{Z}\le\mathbf{u}]Y(s;\gamma_0)\lambda_0(s)ds$ is the compensator of $\mathbf{I}[\mathbf{Z}\le\mathbf{u}]N(s;\gamma_0)$, where $\lambda_0(s)$ is the common hazard function of E_i so that line (2.22) is exactly:

$$\int E\left[Y(s;\gamma_0)\right]E\left[\mathbf{I}[\mathbf{Z}\le\mathbf{u}]Y(s;\gamma_0)\right]\lambda_0(s)ds - \int E\left[\mathbf{I}[\mathbf{Z}\le\mathbf{u}]Y(s;\gamma_0)\right]E\left[Y(s;\gamma_0)\right]\lambda_0(s)ds = 0.$$

Then, it follows from a Taylor series expansion and the expansion of $n^{\frac{1}{2}}(\widetilde{\gamma}-\gamma_0)$,

$$(2.18) = \sqrt{n}(\widetilde{\gamma}-\gamma_0)^\intercal\mathbf{A} + o_p(1) = n^{-\frac{1}{2}}\sum_{i=1}^n \mathbf{A}^\intercal\mathcal{U}_{\gamma i} + o_p(1),$$

where $\mathbf{A} = \int \pi(\mathbf{z}_m,t;\gamma_0)\dot{\mu}(\mathbf{z}_m,dt,\gamma_0) + \int \dot{\pi}(\mathbf{z}_m,t,\gamma_0)\mu(\mathbf{u},dt;\gamma_0) - \int \pi(\mathbf{u},t;\gamma_0)\dot{\mu}(\mathbf{z}_m,dt,\gamma_0) - \int \dot{\pi}(\mathbf{u},t,\gamma_0)\mu(\mathbf{z}_m,dt;\gamma_0)$, $\dot{\mu}(\mathbf{z},t,\gamma) = \partial\mu(\mathbf{z},t,\gamma)/\partial\gamma$ and $\dot{\pi}(\mathbf{z},t,\gamma) = \partial\pi(\mathbf{z},t,\gamma)/\partial\gamma$.

Putting all the aforementioned expansions together, we have $\sqrt{n}\,\widehat{\mathbb{W}}_R(\mathbf{u};\widetilde{\gamma})$ asymptotically equivalent to

$$\begin{aligned}\bar{\mathbb{W}}(\mathbf{u}) = \sqrt{n}\int \Big\{&\pi(\mathbf{z}_m,t;\gamma_0)\widehat{\mathbb{W}}_N(\mathbf{u},dt;\gamma_0) - \pi(\mathbf{u},t;\gamma_0)\widehat{\mathbb{W}}_N(\mathbf{z}_m,dt;\gamma_0)\\ &+\widehat{\mathbb{W}}_\pi(\mathbf{z}_m,t;\gamma_0)\mu(\mathbf{u},dt;\gamma_0) - \widehat{\mathbb{W}}_\pi(\mathbf{u},t;\gamma_0)\mu(\mathbf{z}_m,dt;\gamma_0)\Big\} + n^{-\frac{1}{2}}\sum_{i=1}^n \mathbf{A}^\intercal\mathcal{U}_{\gamma i}.\end{aligned}$$

It then follows from another application of FCLT that $\bar{\mathbb{W}}(\mathbf{u})$ converges weakly to a zero mean Gaussian process $\mathbb{W}(\mathbf{u})$. This together with the smoothness of K, implies that $n\widehat{Q}(\widetilde{\gamma}, \rho)$ converges weakly as a process to

$$\mathcal{Q}(\rho) = \int \int K(\mathbf{u}, \mathbf{v}; \rho) d\mathbb{W}_R(\mathbf{u}) d\mathbb{W}_R(\mathbf{v}).$$

Pathway Selection and Aggregation using Multiple Kernel Learning for Risk Prediction

Jennifer A. Sinnott and Tianxi Cai

Department of Biostatistics

Harvard School of Public Health

Abstract

Attempts to predict risk using high dimensional genomic data can be made difficult by the large number of features and the potential complexity of the relationship between features and the outcome. Integrating prior biological knowledge into risk prediction with such data by grouping genomic features into pathways and networks reduces the dimensionality of the problem and could improve models by making them more biologically grounded and interpretable. Pathways could have complex signals, so our approach to model pathway effects should allow for this complexity. The kernel machine framework has been proposed to model pathway effects because it allows for nonlinear relationships within pathways; it has been used to make predictions for various types of outcomes from individual pathways (Scholkopf and Smola, 2002; Liu et al., 2007, 2008; Li and Luan, 2003; Cai et al., 2011; Liu et al., 2010). When multiple pathways are under consideration, we propose a multiple kernel learning approach to select important pathways and efficiently combine information across pathways. We derive our approach for a general survival modeling framework with a convex objective function, and illustrate its application under the Cox proportional hazards and accelerated failure time (AFT) models. Numerical studies with the AFT model demonstrate that this approach performs well in predicting risk. The methods are illustrated with an application to breast cancer data.

3.1 Introduction

Studies relating disease outcomes to large-scale genomic data are rich resources for improving both our understanding of the progression of disease and our ability to predict patient prognosis. However, such studies usually have many more genomic covariates than study participants which can make it hard to differentiate true biological relationships from noise and false positive associations. Many different approaches for solving this problem have been proposed. Some approaches focus on screening out unimportant markers (e.g., Zhao and Li, 2011) while others build models with penalized coefficients (e.g., Zhang and Lu, 2007). Such single-gene-based approaches, which allow genes to enter or leave a model individually, produce lists of important genes which can be difficult to interpret or replicate. An alternative strategy is to first group genes into biologically relevant sets such as pathways or networks, and relate the overall pathway effects to survival. Pathway-based methods can improve interpretability because the pathways are defined by known or hypothesized functions, which can facilitate generation of mechanistic hypotheses. Moreover, they reduce dimensionality because the number of pathways is generally much smaller than the number of genes. A number of test-based pathway methods are designed to identify important pathways for follow-up (e.g., Goeman et al., 2005); other prediction-based methods are designed to build multivariate regression models (e.g., Wang et al., 2009).

With a few exceptions (e.g., Wei and Li, 2007; Luan and Li, 2008), most existing methods focus primarily on linear effects; however, it is likely that biological pathways have more complex, nonlinear signals due to the presence of feedback loops, signal cascades and/or gene-gene interactions. Kernel machine (KM) modeling is an attractive tool for quantifying complex pathway effects because it allows for non-linear effects without explicitly specifying the forms of those effects. When there is a single pathway of interest, methods have been developed to model pathway effects for non-censored outcomes (Liu et al., 2007, 2008); for censored survival outcomes, KM methods have been proposed for the Cox proportional hazards model (Li and Luan, 2003; Cai et al., 2011) and for the accelerated failure time (AFT) model (Liu et al., 2010; Sinnott and Cai, unpublished). When there are multiple pathways under investigation and some of the pathways may not relate to the event outcome of interest, it is important to select informative pathways for prediction, and efficiently estimate their joint effects. Here, we propose using the KM framework to leverage pathway structures via multiple kernel learning (MKL) (Bach et al., 2004; Lanckriet et al., 2004a) to incorporate multiple pathways. MKL has been proposed in the literature to aggregate information from various types of

genetic data (Lanckriet et al., 2004b; Sonnenburg et al., 2006) for analyzing non-censored outcomes.

In this paper, we propose the use of the MKL framework under various survival models to construct accurate risk prediction rules for censored survival outcomes. In Section 3.2, we derive our methods for a general survival modeling framework with a convex objective function L_0 and propose the use of quadratic approximation for easy computation. In Section 3.3, we illustrate these methods for (i) the Cox model, with L_0 being the log partial likelihood; and (ii) the AFT model, with L_0 being the smoothed Gehan objective function. We demonstrate the procedure using the AFT model in simulation (Section 3.4) and in a data analysis application to a breast cancer gene expression data set (Section 3.5). Concluding remarks are in Section 3.6.

3.2 Approach with a General Objective Function

Let T denote the survival time, $\mathbf{D}$ the $p_{\mathbf{D}} \times 1$ vector of clinical covariates, and $\mathbf{Z}$ the $p_{\mathbf{Z}} \times 1$ vector of genomic measurements. Due to censoring, we observe $X = \min\{T, C\}$ and $\Delta = \mathbf{I}\{T \leq C\}$, where C is a censoring time that is assumed to be independent of T given $\mathbf{W} = (\mathbf{D}^\intercal, \mathbf{Z}^\intercal)^\intercal$. The observed data consist of n independent and identically distributed (iid) random vectors, $\mathcal{O} = \{(X_i, \Delta_i, \mathbf{W}_i^\intercal) : i = 1, \ldots, n\}$. We assume that the genomic covariates $\mathbf{Z}$ are grouped into M pathways with the m^{th} pathway denoted by $\mathbf{Z}_m$; we further assume that these pathways are disjoint.

3.2.1 Kernel Machine Modeling

Suppose the overall effect of $\mathbf{W}$ on T can be summarized as

$$\mu(\boldsymbol{\theta}_0, \mathbf{h} = (h_1, \ldots, h_M)) = \boldsymbol{\theta}_0^\intercal \mathbf{D} + h_1(\mathbf{Z}_1) + \cdots + h_M(\mathbf{Z}_M) \tag{3.1}$$

where $\boldsymbol{\theta}_0 \in \mathbb{R}^{p_{\mathbf{D}}}$, and the $h_m(\cdot)$ are centered, smooth functions quantifying the pathway effect for the m^{th} pathway for $m = 1, \ldots, M$. We assume that each $h_m \in \mathcal{H}_{K_m}$, the Hilbert space generated by some positive definite kernel $K_m(\cdot, \cdot; \rho_m)$. A kernel K is a measure of similarity between two vectors of genomic measurements, and may depend on a possibly unknown scaling parameter ρ. Different choices of kernel will yield different collections of possible functions $h(\cdot)$. For example, the *linear kernel* $K(\mathbf{z}_1, \mathbf{z}_2) = \mathbf{z}_1^\intercal \mathbf{z}_2$ leads to $h(\mathbf{z}) = \boldsymbol{\beta}^\intercal \mathbf{z}$, a linear function of the covariates. To allow for complex non-linear effects, one may consider the *Gaussian kernel*, defined by $K(\mathbf{z}_1, \mathbf{z}_2; \rho) = \exp\{-\|\mathbf{z}_1 - \mathbf{z}_2\|^2/\rho\}$; the resulting function space $\mathcal{H}_K$

is generated by the radial basis functions. For notational simplicity, we will suppress ρ from K, but will discuss selection of ρ when needed.

Suppose a proper convex objective function such as a partial likelihood function, denoted by $L_0(\boldsymbol{\theta}_0, \mathbf{h})$, exists for estimating the unknown parameters. To incorporate potentially high dimensionality in $\mathbf{h}$, one may obtain estimators for $(\boldsymbol{\theta}_0, \mathbf{h})$ by minimizing a penalized objective function

$$L(\boldsymbol{\theta}_0, \mathbf{h}) = L_0(\boldsymbol{\theta}_0, \mathbf{h}) + \frac{c^2}{2} \sum_{m=1}^{M} \|h_m\|^2_{\mathcal{H}_{K_m}}$$

where $\|h\|_{\mathcal{H}_K}$ is the norm of h in $\mathcal{H}_K$ and c is a tuning parameter. The norm of h quantifies the smoothness of h, with smaller values reflecting a smoother function. To further leverage pathway structure and enable pathway selection, we take a MKL approach and further penalize $L(\boldsymbol{\theta}_0, \mathbf{h})$ by the sum of the norms of the h_m:

$$L_{MKL}(\boldsymbol{\theta}_0, \mathbf{h}) = L(\boldsymbol{\theta}_0, \mathbf{h}) + \sum_{m=1}^{M} \lambda_m \|h_m\|_{\mathcal{H}_{K_m}}.$$

Each λ_m is a tuning parameter associated with the m^{th} pathway. This penalty has the effect of setting some pathway effects to 0. When the spaces $\mathcal{H}_{K_m}$ are linearly dependent, the individual pathway effects $h_1, \ldots, h_M$ may not be identifiable even when the overall effect $h_\cdot(\mathbf{z}) = h_1(\mathbf{z}) + \cdots + h_M(\mathbf{z})$ is. Because we want pathway-level information, we will assume the spaces $\mathcal{H}_{K_m}$ are linearly independent, in which case the effect $h_\cdot$ has a unique decomposition $h_\cdot = h_1 + \cdots + h_M$ for $h_m \in \mathcal{H}_{K_m}$. This is a reasonable assumption when the pathways are disjoint as we assume here, but may not be reasonable in a setting with overlapping pathways.

In Lemma 1 in the Appendix, we mimic the proof of the representer theorem (Kimeldorf and Wahba, 1970) in Scholkopf and Smola (2002) to argue that for any $\boldsymbol{\theta}_0$, the minimizers $(\widehat{h}_1, \ldots, \widehat{h}_m)$ of $L_{MKL}(\boldsymbol{\theta}_0, \mathbf{h})$ take a dual form $\widehat{h}_m(\mathbf{z}) = \sum_{i=1}^{n} \alpha_{mi} K_m(\mathbf{z}, \mathbf{z}_i)$. Using the dual representation, the vector $\widehat{\mathbf{h}}_m = (\widehat{h}_m(\mathbf{z}_1), \ldots, \widehat{h}_m(\mathbf{z}_n))^\intercal = \mathbb{K}_m \boldsymbol{\alpha}_m$, where $\mathbb{K}_m$ is the matrix with $(i,j)^{th}$ entry $K_m(\mathbf{z}_i, \mathbf{z}_j)$, and $\|\widehat{h}_m\|^2_{\mathcal{H}_{K_m}} = \alpha_m^\intercal \mathbb{K}_m \alpha_m$, so we may rewrite L_{MKL} as a function of $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1^\intercal, \ldots, \boldsymbol{\alpha}_M^\intercal)^\intercal$ and $\boldsymbol{\theta}_0$:

$$L_{MKL}(\boldsymbol{\theta}_0, \boldsymbol{\alpha}) = L(\boldsymbol{\theta}_0, \boldsymbol{\alpha}) + \sum_{m=1}^{M} \lambda_m \sqrt{\boldsymbol{\alpha}_m^\intercal \mathbb{K}_m \boldsymbol{\alpha}_m}. \tag{3.2}$$

We may further rewrite this by employing a spectral decomposition for $\mathbb{K}_m$. If we let the eigenvalues and associated eigenvectors of $\mathbb{K}_m$ be $\widehat{\eta}_{ml}$ and $\widehat{\boldsymbol{\zeta}}_{ml}$ respectively, for $l = 1, \ldots, n$, where we assume that $\widehat{\eta}_{m1} \geq \cdots \geq \widehat{\eta}_{mn}$ and that the $\widehat{\boldsymbol{\zeta}}_{ml}$ are orthogonal with norm 1, then we may write $\mathbb{K}_m = \widetilde{\mathbb{B}}_m \widetilde{\mathbb{B}}_m^\intercal$, where $\widetilde{\mathbb{B}}_m = \left(\sqrt{\widehat{\eta}_{m1}}\widehat{\boldsymbol{\zeta}}_{m1} \cdots \sqrt{\widehat{\eta}_{mn}}\widehat{\boldsymbol{\zeta}}_{mn}\right)$. Then $\widehat{\mathbf{h}}_m = \widetilde{\mathbb{B}}_m \boldsymbol{\theta}_m$, where $\boldsymbol{\theta}_m = \widetilde{\mathbb{B}}_m^\intercal \boldsymbol{\alpha}_m$, and letting $\boldsymbol{\theta} = (\boldsymbol{\theta}_0^\intercal, \boldsymbol{\theta}_1^\intercal, \ldots, \boldsymbol{\theta}_M^\intercal)^\intercal$,

$L_{MKL}(\boldsymbol{\theta}_0, \boldsymbol{\alpha})$ becomes:

$$L_{MKL}(\boldsymbol{\theta}) = L(\boldsymbol{\theta}) + \sum_{m=1}^{M} \lambda_m \sqrt{\boldsymbol{\theta}_m^\intercal \boldsymbol{\theta}_m}. \tag{3.3}$$

3.2.2 Kernel PCA

One motivation for using a KM approach is to gain power to detect signal in the data; however, each $\boldsymbol{\alpha}_m$ in (3.2) has n components, so by using KMs, we have introduced a large number of parameters, which may cause a loss in power. Parametrizing the problem in terms of the $\boldsymbol{\theta}_m$ does not fix this problem because each $\boldsymbol{\theta}_m$ also has n components, but this parametrization relates to the primal representation of h_m in the Hilbert space $\mathcal{H}_{K_m}$, and leads to a natural strategy for dimension reduction through kernel PCA (Mika et al., 1999; Schölkopf et al., 1998). Specifically, each space $\mathcal{H}_K$ has a (possibly infinite) basis of orthonormal eigenfunctions $\zeta_l(\cdot)$ and eigenvalues η_l of the integral transform $T : \mathcal{H}_K \to \mathcal{H}_K$ defined by $[Tf](\mathbf{z}) = \int K(\mathbf{z}, \mathbf{z}')f(\mathbf{z}')d\mathbf{z}'$, where $\eta_1 \geq \eta_2 \geq \cdots$. Any function $h \in \mathcal{H}_K$ may be written in its primal representation, $h(\mathbf{z}) = \sum_{l=1}^{\infty} \theta_l \sqrt{\eta_l}\zeta_l(\mathbf{z})$, and if the eigenvalues decay quickly, can be well-approximated by a truncated sum $h_{r_0}(\mathbf{z}) = \sum_{l=1}^{r_0} \theta_l \sqrt{\eta_l}\zeta_l(\mathbf{z})$ for some sufficiently large r_0. Moreover, it has been shown that the eigenvalues and eigenvectors obtained based on $\mathbb{K}$ can be used to consistently estimate the underlying true eigenvalues and eigenfunctions of $\mathcal{H}_K$ (Koltchinskii and Giné, 2000; Braun, 2005). Thus, by estimating the coefficients $\boldsymbol{\theta}_m$, we are estimating h in its approximate primal form, and when the eigenvalues decay quickly, we may not lose much information by estimating h using only the first r_m eigenvectors, where r_m is the smallest number for which $\sum_{i=1}^{r_m} \widehat{\eta}_{mi} / \sum_{i=1}^{n} \widehat{\eta}_{mi} \geq \mathfrak{p}$ for a prespecified fraction $\mathfrak{p}$. Ideally, the included eigenvectors encode aspects of maximal variability in the data, while the excluded eigenvectors capture noise. This approximation is called the kernel PCA approximation, and it can greatly reduce the number of unknown parameters being estimated because r_m is frequently significantly smaller than n.

Writing the truncated matrices $\widetilde{\mathbb{B}}_{mr_m} = \left(\sqrt{\widehat{\eta}_{m1}}\widehat{\boldsymbol{\zeta}}_{m1} \cdots \sqrt{\widehat{\eta}_{mr_m}}\widehat{\boldsymbol{\zeta}}_{mr_m}\right)$, we replace $\mathbb{K}_m$ by $\mathbb{K}_{mr_m} = \widetilde{\mathbb{B}}_{mr_m}\widetilde{\mathbb{B}}_{mr_m}^\intercal$ in (3.2). For notational simplicity, however, we will continue to write $\boldsymbol{\alpha}_m$ and $\boldsymbol{\theta}_m$ for the coefficients associated with these approximated matrices; since we recover $\widetilde{\mathbb{B}}_m$ and $\mathbb{K}_m$ by taking $\mathfrak{p} = 1$, the kernel PCA formulation is in fact simply more general. Thus, we proceed with $L_{MKL}(\boldsymbol{\theta})$ in (3.3) as our objective function, keeping in mind that we may be using a kernel PCA approximation.

3.2.3 Least Squares Approximation

The penalty in (3.3) is equivalent to the group lasso penalty; the equivalence between MKL and the group lasso has been noted in Bach (2008). Methods for fitting a model with this penalty have been worked out for linear and logistic regression (Yuan and Lin, 2006; Meier et al., 2008). Rather than developing the machinery to minimize (3.3) for specific functions $L(\boldsymbol{\theta})$, we propose to approximate $L(\boldsymbol{\theta})$ via a quadratic approximation similar to those proposed in Wang and Leng (2007) and Zhang and Lu (2007). In our setting, we first consider the Taylor series expansion of $L(\boldsymbol{\theta})$ about its minimizer $\widetilde{\boldsymbol{\theta}}$:

$$L(\boldsymbol{\theta}) \approx L(\widetilde{\boldsymbol{\theta}}) + \dot{L}(\widetilde{\boldsymbol{\theta}})^{\intercal}(\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}) + \frac{1}{2}(\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}})^{\intercal}\ddot{L}(\widetilde{\boldsymbol{\theta}})(\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}).$$

where $\dot{L}(\cdot)$ and $\ddot{L}(\cdot)$ are the first- and second-order derivatives with respect to $\boldsymbol{\theta}$. The term $L(\widetilde{\boldsymbol{\theta}})$ is constant, and $\dot{L}(\widetilde{\boldsymbol{\theta}}) = 0$ since $\widetilde{\boldsymbol{\theta}}$ minimizes $L(\boldsymbol{\theta})$, so we may approximate the original objective function near $\widetilde{\boldsymbol{\theta}}$ by the quadratic function:

$$\frac{1}{2}(\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}})^{\intercal}\ddot{L}(\widetilde{\boldsymbol{\theta}})(\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}).$$

Thus, near $\widetilde{\boldsymbol{\theta}}$, minimizing $L_{MKL}(\boldsymbol{\theta})$ is equivalent to minimizing

$$Q(\boldsymbol{\theta}) = \frac{1}{2}(\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}})^{\intercal}\ddot{L}(\widetilde{\boldsymbol{\theta}})(\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}) + \sum_{m=1}^{M} \lambda_m \sqrt{\boldsymbol{\theta}_m^{\intercal}\boldsymbol{\theta}_m}.$$

This minimization may be done using existing software for the group lasso for linear regression applied to pseudodata. Specifically, by taking the squareroot matrix of $\ddot{L}(\widetilde{\boldsymbol{\theta}})$, say $\ddot{L}(\widetilde{\boldsymbol{\theta}}) = \widetilde{X}_{pseudo}^{\intercal}\widetilde{X}_{pseudo}$, and letting $\widetilde{Y}_{pseudo} = \widetilde{X}_{pseudo}\widetilde{\boldsymbol{\theta}}$, we have:

$$Q(\boldsymbol{\theta}) = \frac{1}{2}(\widetilde{X}_{pseudo}\boldsymbol{\theta} - \widetilde{Y}_{pseudo})^{\intercal}(\widetilde{X}_{pseudo}\boldsymbol{\theta} - \widetilde{Y}_{pseudo}) + \sum_{m=1}^{M} \lambda_m \sqrt{\boldsymbol{\theta}_m^{\intercal}\boldsymbol{\theta}_m}.$$

a standard least squares formulation.

Minimizing $Q(\boldsymbol{\theta})$ will result in an estimator $\widehat{\boldsymbol{\theta}}$ where some pathways may have had all coefficients set to 0. We could now use $\widehat{\boldsymbol{\theta}}$ as our estimate of $\boldsymbol{\theta}$, or we could iterate the procedure, restricting the data to the retained pathways to re-estimate $\widetilde{\boldsymbol{\theta}}$ the minimizer of $L(\boldsymbol{\theta})$, and minimizing $Q(\boldsymbol{\theta})$ centered at this $\widetilde{\boldsymbol{\theta}}$. We could repeat until the collection of pathways has stabilized. We compare the results of this iterative procedure to the results of using the first estimated $\widehat{\boldsymbol{\theta}}$.

3.2.4 Kernel Selection and Tuning

To implement the procedure, we need to select a kernel K_m for each pathway, select a tuning parameter ρ_m if the kernel requires it, select the ridge-type penalty parameter c in $L(\boldsymbol{\theta})$, and select the group lasso penalty

parameters λ_m.

To select a kernel, researchers may use subject matter knowledge to decide on which kernel best captures similarity in their data. They may also be guided by what scope of models they wish to consider (e.g., linear or nonlinear). If several kernels are of interest, it may be desirable to have a data-driven approach to kernel selection. For both the Cox and AFT models considered later on, KM pathway tests have been developed to test the null hypothesis $H_0 : \mathbf{h}_m(\cdot) = 0$ in a marginal model relating survival to $\boldsymbol{\theta}_0^{\intercal}\mathbf{D} + h_m(\mathbf{Z}_m)$ (Cai et al., 2011; Sinnott and Cai, unpublished). In these settings, one may perform the marginal test for different kernels of interest and use the kernel which yields the smallest p-value. We may also use these marginal KM pathway tests to select tuning parameters ρ_m. We can test $H_0 : h_m(\cdot) = 0$ in the marginal pathway model for different ρ_m, and use the value of ρ_m that corresponds to the smallest p-value.

To choose the ridge-type tuning parameter c in $L(\boldsymbol{\theta})$, we propose using the value of c which minimizes a modified Bayesian Information Criterion (BIC) type penalty defined by:

$$BIC(\boldsymbol{\theta}(c)) = -2\log\widehat{\text{lik}}(\boldsymbol{\theta}(c)) + \widehat{\text{df}}(c)n^{0.1},$$

where $\widehat{\text{lik}}(\boldsymbol{\theta})$ is an estimate of the likelihood at $\boldsymbol{\theta}$, whose form will depend on the model, and $\widehat{\text{df}}(c)$ is an estimate of the degrees of freedom, given by $\widehat{\text{df}}(c) = \text{tr}R(c)$,

$$R(c) = \widetilde{\mathbb{W}}^{\intercal}\widetilde{\mathbb{W}}\left(\widetilde{\mathbb{W}}^{\intercal}\widetilde{\mathbb{W}} + n\frac{c^2}{2}I_{p_{\widetilde{\mathbf{W}}}}\right)^{-1}$$

where $\widetilde{\mathbb{W}}$ is the $n \times p_{\widetilde{\mathbf{W}}}$ matrix with i^{th} row $\widetilde{\mathbf{W}}_i = (\mathbf{D}_i^{\intercal}, \widetilde{\mathbf{B}}_{1i}^{\intercal}, \ldots, \widetilde{\mathbf{B}}_{Mi}^{\intercal})^{\intercal}$ for $\widetilde{\mathbf{B}}_{mi}$ the i^{th} row of $\widetilde{\mathbb{B}}_{mr_m}$, and $p_{\widetilde{\mathbf{W}}} = p_{\mathbf{D}} + \sum_{m=1}^{M} r_m$. Here, $n^{0.1}$ is chosen to replace $\log(n)$ to improve the prediction performance in finite sample since $\log(n) \gg n^{0.1}$ and the standard BIC criterion tends to bias the fit too far towards the null.

To choose the group lasso-type tuning parameters λ_m, we propose a standard adaptive group lasso penalty, $\lambda_m = \lambda/\|\widetilde{\boldsymbol{\theta}}_m(c)\|$, where $\|\boldsymbol{\theta}\| = \sqrt{\boldsymbol{\theta}^{\intercal}\boldsymbol{\theta}}$. To select λ, we again use a modified BIC penalty defined by:

$$BIC(\boldsymbol{\theta}(\lambda)) = -2\log\widehat{\text{lik}}(\boldsymbol{\theta}(\lambda)) + \tilde{\text{df}}(\boldsymbol{\theta}(\lambda))n^{0.1}.$$

Here, $\tilde{\text{df}}(\theta(\lambda))$ is an estimate of the degrees of freedom similar to that suggested in the group lasso literature:

$$\tilde{\text{df}}(\boldsymbol{\theta}) = \sum_{m=1}^{M} I\{\|\boldsymbol{\theta}_m\| > 0\} + \sum_{m=1}^{M} \frac{\|\boldsymbol{\theta}_m\|}{\|\widetilde{\boldsymbol{\theta}}_m(c)\|}(\widehat{\text{df}}_m(c) - 1).$$

and we once again replace $\log(n)$ by $n^{0.1}$ since the standard BIC criterion tends to set too many pathways to be non-informative. The quantity $\widehat{\text{df}}_m(c)$ is an estimate of the degrees of freedom in the m^{th} pathway, found by summing the diagonal entries of $R(c)$ associated with the m^{th} pathway.

3.2.5 Pathway Screening

When the number of pathways is large, the number of components in $\boldsymbol{\theta}$ can become quite large even after using kernel PCA approximations, and the problem can become computationally difficult. When it is reasonable to hypothesize that the number of biological pathways related to survival is much smaller than the total number of pathways under consideration and when there is a marginal pathway test available for the model of interest, it may be reasonable to perform a preliminary screening test, where each pathway is tested marginally and only retained if the pathway p-value passes some FWER or FDR threshold.

3.3 Examples

3.3.1 Cox Model

The Cox PH KM model with M pathways assumes:

$$\lambda_i(t) = \lambda_0(t)\exp\{\boldsymbol{\theta}_0^{\mathsf{T}}\mathbf{D}_i + h_1(\mathbf{Z}_{1i}) + \cdots + h_M(\mathbf{Z}_{Mi})\}, i = 1, \ldots, n$$

where $\lambda_i(t)$ is the hazard that person i has an event at time t given their covariates $\mathbf{W}_i$ and $\lambda_0(t)$ is a common baseline hazard function. Defining the usual counting and at risk processes $N_i(t) = \Delta_i I\{X_i \leq t\}$ and $Y_i(t) = I\{X_i \geq t\}$, we can let $L(\boldsymbol{\theta})$ be the penalized log partial likelihood function:

$$L(\boldsymbol{\theta}) = \sum_{i=1}^{n}\int\left[\boldsymbol{\theta}^{\mathsf{T}}\widetilde{\mathbf{W}}_i - \log\left\{S^{(0)}(\boldsymbol{\theta}, s)\right\}\right]dN_i(s) + \frac{c^2}{2}\sum_{m=1}^{M}\boldsymbol{\theta}_m^{\mathsf{T}}\boldsymbol{\theta}_m.$$

where $S^{(k)}(\boldsymbol{\theta}, s) = \sum_{l=1}^{n} Y_l(s)\exp\{\boldsymbol{\theta}^{\mathsf{T}}\widetilde{\mathbf{W}}_l\}\widetilde{\mathbf{W}}_l^{\otimes k}$. Then:

$$\ddot{L}(\boldsymbol{\theta}) = \sum_{i=1}^{n}\int\left\{\frac{S^{(2)}(\boldsymbol{\theta}, s)}{S^{(0)}(\boldsymbol{\theta}, s)} - \frac{S^{(1)}(\boldsymbol{\theta}, s)^{\otimes 2}}{S^{(0)}(\boldsymbol{\theta}, s)^2}\right\}dN_i(s) + c^2 G$$

where G is the diagonal matrix whose first $p_{\mathbf{D}}$ diagonal entries are 0 and whose remaining $r_{\cdot}$ diagonal entries are 1. KM testing for individual pathways can be done using the method developed in Cai et al. (2011). We can use $L(\boldsymbol{\theta})$ as the likelihood in BIC calculations.

3.3.2 AFT model

The AFT-KM model with M pathways assumes:

$$\log T_i = \boldsymbol{\theta}_0^{\mathsf{T}}\mathbf{D}_i + h_1(\mathbf{Z}_{1i}) + \cdots + h_M(\mathbf{Z}_{Mi}) + E_i,\ i = 1, \ldots, n \tag{3.4}$$

where E_i is an iid error term independent of $\mathbf{W}_i = (\mathbf{D}_i,^\intercal \mathbf{Z}_i^\intercal)^\intercal$ with completely unspecified distribution. For this model, we can let $L(\boldsymbol{\theta})$ be the penalized Gehan objective function:

$$L(\boldsymbol{\theta}) = n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} \Delta_i |\widetilde{e}_j(\boldsymbol{\theta}) - \widetilde{e}_i(\boldsymbol{\theta})|_+ + \frac{c^2}{2} \sum_{m=1}^{M} \boldsymbol{\theta}_m^\intercal \boldsymbol{\theta}_m \tag{3.5}$$

where $\widetilde{e}_i(\boldsymbol{\theta}) = \log X_i - \boldsymbol{\theta}^\intercal \widetilde{\mathbf{W}}_i$. Unfortunately, this objective function is only once differentiable, so our procedure does not directly apply. To remedy this, we perform a smoothing step, following reasoning similar to that in Brown and Wang (2007). Specifically, the gradient of $L(\boldsymbol{\theta})$ is:

$$\dot{L}(\boldsymbol{\theta}) = n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} \Delta_i (\widetilde{\mathbf{W}}_i - \widetilde{\mathbf{W}}_j) \mathbf{I}\left\{\widetilde{e}_j(\boldsymbol{\theta}) - \widetilde{e}_i(\boldsymbol{\theta}) > 0\right\} + c^2 G\boldsymbol{\theta} \tag{3.6}$$

where G is the diagonal matrix whose first $p_{\mathbf{D}}$ diagonal entries are 0 and whose remaining $r_{\cdot}$ diagonal entries are 1. The function $\dot{L}$ has jumps because of the indicator function $\mathbf{I}\{\cdot\}$, so to smooth $\dot{L}$ we could replace $\mathbf{I}\{\widetilde{e}_j(\boldsymbol{\theta}) - \widetilde{e}_i(\boldsymbol{\theta}) > 0\}$ by $\Phi\left(\frac{e_j(\boldsymbol{\theta}) - e_i(\boldsymbol{\theta})}{\sigma_n}\right)$ where Φ is some continuous cdf and σ_n is a bandwidth parameter. Here, we will take Φ to be the standard normal cdf, and use as a bandwidth $\sigma_n = \text{s.d.}\{e_j(\widetilde{\boldsymbol{\theta}})\} \times n^{-\frac{1}{3}}$. We choose this bandwidth with under-smoothing to eliminate the potential bias induced by smoothing (van der Vaart, 1994). Our smoothed version of $\dot{L}$ is $\dot{L}_{sm}$:

$$\dot{L}_{sm}(\boldsymbol{\theta}) = n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} \Delta_i (\widetilde{\mathbf{W}}_i - \widetilde{\mathbf{W}}_j) \Phi\left(\frac{e_j(\boldsymbol{\theta}) - e_i(\boldsymbol{\theta})}{\sigma_n}\right) + c^2 G\boldsymbol{\theta} \tag{3.7}$$

We can now take a further derivative:

$$\ddot{L}_{sm}(\boldsymbol{\theta}) = \frac{\partial}{\partial(\boldsymbol{\theta})^\intercal} \dot{L}_{sm}(\boldsymbol{\theta}) = n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\Delta_i}{\sigma_n} \phi\left(\frac{\widetilde{e}_j(\boldsymbol{\theta}) - \widetilde{e}_i(\boldsymbol{\theta})}{\sigma_n}\right) (\widetilde{\mathbf{W}}_i - \widetilde{\mathbf{W}}_j)(\widetilde{\mathbf{W}}_i - \widetilde{\mathbf{W}}_j)^\intercal + c^2 G \tag{3.8}$$

Finally, we can check that $\dot{L}_{sm}(\boldsymbol{\theta})$ is the gradient of the smoothed Gehan objective function defined by:

$$L_{sm}(\boldsymbol{\theta}) = n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} \Delta_i \left[(e_j(\boldsymbol{\theta}) - e_i(\boldsymbol{\theta}))\Phi\left(\frac{e_j(\boldsymbol{\theta}) - e_i(\boldsymbol{\theta})}{\sigma_n}\right) + \sigma\phi\left(\frac{e_j(\boldsymbol{\theta}) - e_i(\boldsymbol{\theta})}{\sigma_n}\right)\right] + \frac{c^2}{2} \sum_{m=1}^{M} \boldsymbol{\theta}_m^\intercal \boldsymbol{\theta}_m$$

KM testing for individual pathways can be done using the method developed in Sinnott and Cai (unpublished). We can use the log sieve likelihood estimate (Zeng and Lin, 2007) in place of the likelihood in BIC calculations.

3.4 Simulation Studies

To assess the performance of the estimation procedure, we conducted simulation studies. We considered settings with $M = 10$ and $M = 20$ total pathways, with each pathway of size $P = 5$ or $P = 10$. Three

pathways were related to outcome, while the remaining pathways were noise; and when a pathway was related to outcome, all genes in the pathway were involved. The genes were simulated from a multivariate normal with mean 0 and variance 1, and blockwise compound symmetry; we considered within pathway correlation of 0.2 and 0.4, and between pathway correlation of 0 and 0.1. The underlying relationship between the associated pathways and survival was either linear or nonlinear, with all three pathways having the same signal. We considered sample size $N = 200$, % censoring 25%, and for simplicity, no additional covariates.

Table 3.1: Simulated data with 3 related pathways with all genes involved in a linear signal, remaining pathways unrelated. Shown are results from 10 and 20 paths of size 5 and 10, for different levels of between- and within-pathway correlation, for linear kernel (L) and Gaussian kernel (G). Compared are the C-statistic for the AFT lasso (C lasso), the initial KM fit with ridge penalty (C ridge), and the KM fit after one iteration of group selection (C MKL). Also shown is the C from the oracle fit (C oracle), as well as the number of genes for the AFT lasso (N genes lasso) and the KM fit (N genes MKL).

Path Size	5															
N paths	10								20							
bt Correlation	0				0.1				0				0.1			
wi Correlation	0.2		0.4		0.2		0.4		0.2		0.4		0.2		0.4	
Kernel	L	G	L	G	L	G	L	G	L	G	L	G	L	G	L	G
C lasso	71	72	76	76	77	77	79	80	65	65	72	72	73	73	79	78
C ridge	71	71	75	74	76	76	78	77	63	64	69	70	70	72	74	75
C MKL	73	73	77	76	78	78	80	79	66	66	73	73	74	75	79	79
C oracle	75	74	78	77	79	78	81	79	71	70	78	77	77	76	81	80
N genes lasso	28.1	25.8	25.3	24.7	24	23.1	22.6	21.9	60.8	61.4	54.4	57.8	40.7	38.9	38.6	33
N genes MKL	33.6	28.7	32.8	32.8	30.7	27	31.1	31.4	87.7	69.7	87.9	80.4	72	48.6	81.2	61.5
Path Size	10															
C lasso	75	76	83	82	84	84	86	86	70	70	81	80	84	85	88	88
C ridge	73	79	81	83	81	85	83	85	68	77	74	82	78	84	79	86
C MKL	77	80	84	84	85	86	85	86	68	80	75	85	82	87	83	88
C oracle	79	80	85	85	86	86	86	86	81	82	87	87	87	88	88	88
N genes lasso	58.4	58.4	47.7	50.6	39.9	39	37	35	154.7	154.5	114.7	121.6	62.1	64.8	42.6	43.4
N genes MKL	94.7	57.7	89.1	57.8	86.4	50	84	50	200	124.2	200	150.7	170.6	95	191.2	80

We considered fitting all pathways with linear kernel, which requires no tuning, or with Gaussian kernel, which depends on a tuning parameter ρ selected as the one producing the smallest p-value from the

test in Sinnott and Cai (unpublished). Kernel PCA was used with $\mathfrak{p} = 0.95$. As a gold standard, we fit the "oracle" model, where we restricted the fit to the pathways known to be involved, as well as an AFT model with standard adaptive lasso penalty, which is fit using a quadratic approximation to the likelihood similar to that described above.

Results from simulations with linear signal are shown in Table 3.1. When the pathways are of size 5, the KM method tends to improve slightly over the lasso. The linear and Gaussian kernels tend to perform very similarly. When the pathway size increases to 10, we see a slightly greater amount of improvement of the KM method over the lasso when the Gaussian kernel is used. Results from simulations with nonlinear signal are shown in Table 3.2. Here, the linear KM method performs very similarly to the Lasso as we might expect, while the KM method with Gaussian kernel improves often substantially over the lasso. This suggests that our method may improve risk prediction when the signal is nonlinear.

When we let the procedure iterate until the number of pathways stabilized, the stabilization usually happened after a single iteration and so did not alter the fit. Thus, we recommend using a single iteration, since it is faster and the iteration never improved the C statistic.

Overall, the KM approach with Gaussian kernel performs well. The most notable drawback is that the pathway-based models tend to be larger than the lasso models; in some settings the pathway-based method is selecting all the pathways. We might be able to improve this with a more stringent approach to tuning.

3.5 Example: Breast Cancer Gene Expression Study

Genomic information has already improved our understanding of breast cancer. A number of gene expression signatures have been introduced into clinical practice to better identify cancers with high and low risk of recurrence (Desmedt et al., 2011). Despite these advances, approximately 60% of patients with early-stage breast cancer are given adjuvant therapy in addition to local treatment, while only a small proportion are thought to benefit (Reis-Filho and Pusztai, 2011). Better markers of aggressive disease would help physicians predict which patients could safely avoid adjuvant therapy and its negative side-effects, and which patients should be treated with more aggressive therapy.

One promising hypothesis-driven approach is to investigate the effects of candidate pathways on recurrence-free survival. We collected 32 candidate pathways from the molecular signature database. Some of the pathways overlapped, but we ignored the overlap for the purpose of this analysis.

Table 3.2: Simulated data with 3 related pathways with all genes involved in a nonlinear signal, remaining pathways unrelated. Shown are results from 10 and 20 paths of size 5 and 10, for different levels of between- and within-pathway correlation, for linear kernel (L) and Gaussian kernel (G). Compared are the C-statistic for the AFT lasso (C lasso), the initial KM fit with ridge penalty (C ridge), and the KM fit after one iteration of group selection (C MKL). Also shown is the C from the oracle fit (C oracle), as well as the number of genes for the AFT lasso (N genes lasso) and the KM fit (N genes MKL).

Path Size	5															
N paths	10								20							
bt Correlation	0				0.1				0				0.1			
wi Correlation	0.2		0.4		0.2		0.4		0.2		0.4		0.2		0.4	
Kernel	L	G	L	G	L	G	L	G	L	G	L	G	L	G	L	G
C lasso	59	60	59	59	60	61	60	61	56	55	54	54	57	55	56	55
C ridge	57	69	57	73	58	70	58	74	55	63	53	67	55	68	54	68
C MKL	58	70	59	74	60	71	60	75	56	65	54	69	57	70	56	70
C oracle	61	72	61	76	62	72	63	77	60	70	58	74	61	75	61	75
N genes lasso	18.8	14.5	17.8	18.3	19.4	14.6	18.6	19.2	58	56.7	61	60.9	53.5	53.7	54	54.5
N genes MKL	33.7	34.7	32.7	34.3	35	32.2	36.6	35	88.7	76.7	92.8	77.5	87.2	60.3	90.5	80.8
Path Size	10															
C lasso	56	57	56	57	59	59	58	58	53	54	53	54	55	55	56	56
C ridge	56	60	55	68	57	64	57	70	53	56	53	66	54	61	54	66
C MKL	57	60	57	69	59	62	59	70	53	57	53	68	54	57	55	67
C oracle	61	63	60	70	63	66	62	71	59	61	59	70	61	65	61	70
N genes lasso	61.1	53.1	54.1	49.9	50.8	49.9	51.8	50.8	163.8	167.8	162.5	152.6	126.1	133.1	106.3	114
N genes MKL	96.8	44.8	95	62.4	95.6	40	92.9	63.3	200	106.7	187.5	141.8	200	67.9	199.3	156

To assess the effects of these pathways on breast cancer progression, we applied our AFT KM marginal pathway test with the Gaussian kernel to each of the pathways in a training set of 286 lymph node negative breast cancer patients who received no systemic adjuvant therapy (Wang et al., 2005). We used tumor gene expression assessed on the Affymetrix U133a Gene Chip. A total of 107 deaths or recurrences were observed, with follow-up time ranging between 2 months and 14.3 years (median 7.2 years); 63% of observations were censored. We screened pathways at a family-wise error rate of 5%, and used the test to determine which value of the Gaussian kernel tuning parameter ρ_m to use for each pathway. Nine pathways were retained from this screening. We then used our MKL procedure to fit a model selecting from among these nine pathways; all of the pathways were retained for the final model. We applied the result-

ing model to an independent validation set of 119 lymph node negative patients with no adjuvant therapy, with gene expression assessed on the same chip (Sotiriou et al., 2006). In this validation set, 27 deaths or recurrences were observed, with follow-up time ranging between 2 months and 14.5 years (median 7.7 years); 77% of observations were censored. We assessed the accuracy of each model using the C-statistic and calculated 95% confidence intervals (Uno et al., 2011). Our final model produced a C-statistic in the validation set of 66% (95% CI: 54% - 79%).

For comparison, we fit a Cox model with the lasso penalty to the training data, using the 788 genes in the candidate pathways. We did this instead of the AFT model with the lasso penalty fit in simulations because that procedure becomes computationally burdensome when the number of features becomes very large. The Cox model yielded a C-statistic in the validation data of 64% (95% CI: 52% - 77%), which is very similar to the results of our MKL procedure.

3.6 Discussion

In this paper, we proposed KM based procedures to build a risk prediction model by selecting and combining information from multiple pathways. By taking a pathway based approach, we take advantage of pre-existing biological knowledge and privilege groups of genes believed to work together. By working with the Gaussian kernel, we can capture both linear and nonlinear effects well. When the underlying signal is sparse with only a few genes in a pathway associated with survival, it would also be interesting to extend our proposed procedure to allow for feature selection under the KM framework. In reality, biological pathways overlap, so our assumption that pathways are disjoint will generally not be true in practice. Based on existing literature about overlapping group lasso (Jacob et al., 2009), our method should extend easily to the situation where pathways overlap if we change the norm we use to penalize pathways; the computational methods should not change. We hope to elaborate on this in the future. The model as currently formulated also assumes that genes cannot interact between pathways, but it would be interesting to extend the method to allow this to occur.

3.7 Appendix

Lemma 1. *For fixed $\boldsymbol{\theta}_0$, c, and λ_m, the minimizers $\widehat{h}_m(\mathbf{z})$ of*

$$L_{MKL}(\boldsymbol{\theta}_0, \mathbf{h}) = L(\boldsymbol{\theta}_0, \mathbf{h}) + \frac{c^2}{2} \sum_{m=1}^{M} \|h_m\|^2_{\mathcal{H}_{K_m}} + \sum_{m=1}^{M} \lambda_m \|h_m\|_{\mathcal{H}_{K_m}}$$

take a dual form $\widehat{h}_m(\mathbf{z}) = \sum_{i=1}^{n} \alpha_{mi} K_m(\mathbf{z}, \mathbf{z}_i)$.

Proof. For fixed $\boldsymbol{\theta}_0$, c, and λ_m, let $\mathbf{h}$ be a minimizer of $L_{MKL}(\boldsymbol{\theta}_0, \mathbf{h})$. For each m, let $V_m = \text{span}\{K_m(\mathbf{z}_1, \cdot), \ldots, K_m(\mathbf{z}_n, \cdot)\}$ and decompose $\mathcal{H}_{K_m} = V_m \oplus V_m^{\perp}$. Decompose $h_m = h_{V_m} + h_{V_m^{\perp}}$, $h_{V_m} \in V_m$, $h_{V_m^{\perp}} \in V_m^{\perp}$.

Claim 1: The value of $L(\boldsymbol{\theta}_0, \mathbf{h})$ does not depend on $h_{V_m^{\perp}}$. This is because $L(\boldsymbol{\theta}_0, \mathbf{h})$ depends only on h_m evaluated at the observed data $\mathbf{z}_1, \cdots, \mathbf{z}_n$, and $h_m \in \mathcal{H}_{K_m}$ implies that:

$$\begin{aligned} h_m(\mathbf{z}_j) &= \langle h_m(\cdot), K_m(\mathbf{z}_j, \cdot) \rangle \\ &= \langle h_{V_m}(\cdot) + h_{V_m^{\perp}}(\cdot), K_m(\mathbf{z}_j, \cdot) \rangle \\ &= \langle h_{V_m}(\cdot), K_m(\mathbf{z}_j, \cdot) \rangle + \langle h_{V_m^{\perp}}(\cdot), K_m(\mathbf{z}_j, \cdot) \rangle \\ &= h_{V_m}(\mathbf{z}_j) + 0. \end{aligned}$$

Claim 2: The value of $\|h_m\|^2_{\mathcal{H}_{K_m}}$ is smallest when $h_{V_m^{\perp}}(\cdot) = 0$. This is because by orthogonality:

$$\|h_m\|^2_{\mathcal{H}_{K_m}} = \|h_{V_m}\|^2_{\mathcal{H}_{K_m}} + \|h_{V_m^{\perp}}\|^2_{\mathcal{H}_{K_m}}.$$

Thus the minimizers satisfy $h_m \in V_m$, so that $h_m(\mathbf{z}) = \sum_{i=1}^{n} \alpha_{mi} K_m(\mathbf{z}_i, \mathbf{z})$. □

References

ALBERTS, B. *(2010). Editorial expression of concern. Science* ***330*** *912.*

BACH, F. *(2008). Consistency of the group lasso and multiple kernel learning. The Journal of Machine Learning Research* ***9*** *1179–1225.*

BACH, F., LANCKRIET, G. *and* JORDAN, M. *(2004). Multiple kernel learning, conic duality, and the smo algorithm. In Proceedings of the twenty-first international conference on Machine learning. ACM.*

BEECHAM, G. W., MARTIN, E. R., GILBERT, J. R., HAINES, J. L. *and* PERICAK-VANCE, M. A. *(2010). APOE is not associated with alzheimer disease: a cautionary tale of genotype imputation. Ann Hum Genet* ***74*** *189–94.*

BILIAS, Y., GU, M. *and* YING, Z. *(1997). Towards a general asymptotic theory for cox model with staggered entry. The Annals of Statistics 662–682.*

BRAUN, M. *(2005). Spectral properties of the kernel matrix and their application to kernel methods in machine learning. Ph.D. thesis, PhD thesis, University of Bonn.*

BROWN, B. *and* WANG, Y. *(2007). Induced smoothing for rank regression with censored survival times. Statistics in medicine* ***26*** *828–836.*

BUCKLEY, J. *and* JAMES, I. *(1979). Linear regression with censored data. Biometrika* ***66*** *429.*

CAI, T., TONINI, G. *and* LIN, X. *(2011). Kernel machine approach to testing the significance of multiple genetic markers for risk prediction. Biometrics .*

CARMICHAEL, M. *(2010). The little flaw in the longevity-gene study that could be a big problem.*

URL `http://www.newsweek.com/2010/07/07/the-little-flaw-in-the-longevity-gene-study-that-`

`html`

DAVIES, R. *(1987). Hypothesis testing when a nuisance parameter is present only under the alternative. Biometrika* ***74*** *33–43.*

DESMEDT, C., MICHIELS, S., HAIBE-KAINS, B., LOI, S. *and* SOTIRIOU, C. *(2011). Time to move forward from first-generation prognostic gene signatures in early breast cancer. Breast Cancer Research and Treatment 1–3.*

DEVLIN, B. *and* ROEDER, K. *(1999). Genomic control for association studies. Biometrics* ***55*** *997–1004.*

FALLIN, M. D., SZYMANSKI, M., WANG, R., GHERMAN, A., BASSETT, S. S. *and* AVRAMOPOULOS, D. *(2010). Fine mapping of the chromosome 10q11-q21 linkage region in Alzheimer's disease cases and controls. Neurogenetics* ***11*** *335–48.*

FISHER, R. A. *(1925). Statistical methods for research workers. Oliver and Boyd, Edinburgh, London.*

GASCO, M., SHAMI, S. *and* CROOK, T. *(2002). The p53 pathway in breast cancer. Breast Cancer Research* ***4*** *70–76.*

GOEMAN, J., OOSTING, J., CLETON-JANSEN, A., ANNINGA, J. *and* VAN HOUWELINGEN, H. *(2005). Testing association of a pathway with survival using gene expression data. Bioinformatics* ***21*** *1950–1957.*

HO, L. A. *and* LANGE, E. M. *(2010). Using public control genotype data to increase power and decrease cost of case-control genetic association studies. Hum Genet* ***128*** *597–608.*

HOM, G., GRAHAM, R. R., MODREK, B., TAYLOR, K. E., ORTMANN, W., GARNIER, S., LEE, A. T., CHUNG, S. A., FERREIRA, R. C., PANT, P. V. K., BALLINGER, D. G., KOSOY, R., DEMIRCI, F. Y., KAMBOH, M. I., KAO, A. H., TIAN, C., GUNNARSSON, I., BENGTSSON, A. A., RANTAPAA-DAHLQVIST, S., PETRI, M., MANZI, S., SELDIN, M. F., RONNBLOM, L., SYVANEN, A.-C., CRISWELL, L. A., GREGERSEN, P. K. *and* BEHRENS, T. W. *(2008). Association of systemic lupus erythematosus with C8orf13-BLK and ITGAM-ITGAX. N Engl J Med* ***358*** *900–909.*

HOWIE, B. N., DONNELLY, P. *and* MARCHINI, J. *(2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet* ***5*** *e1000529.*

HUNTER, D. J., KRAFT, P., JACOBS, K. B., COX, D. G., YEAGER, M., HANKINSON, S. E., WACHOLDER, S., WANG, Z., WELCH, R., HUTCHINSON, A., WANG, J., YU, K., CHATTERJEE, N., ORR, N., WILLETT, W. C., COLDITZ, G. A., ZIEGLER, R. G., BERG, C. D., BUYS, S. S., MCCARTY, C. A., FEIGELSON, H. S., CALLE, E. E., THUN, M. J., HAYES, R. B., TUCKER, M., GERHARD, D. S., FRAUMENI, J. F. J., HOOVER,

R. N., THOMAS, G. *and* CHANOCK, S. J. *(2007). A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. Nat Genet* ***39*** *870–874.*

JACOB, L., OBOZINSKI, G. *and* VERT, J. *(2009). Group lasso with overlap and graph lasso. In Proceedings of the 26th Annual International Conference on Machine Learning. ACM.*

JIN, Z., LIN, D., WEI, L. *and* YING, Z. *(2003). Rank-based inference for the accelerated failure time model. Biometrika* ***90*** *341.*

JIN, Z., YING, Z. *and* WEI, L. *(2001). A simple resampling method by perturbing the minimand. Biometrika* ***88*** *381.*

KALBFLEISCH, J., PRENTICE, R. *and* KALBFLEISCH, J. *(1980). The statistical analysis of failure time data, vol. 5. Wiley New York.*

KIMELDORF, G. *and* WAHBA, G. *(1970). A correspondence between bayesian estimation on stochastic processes and smoothing by splines. The Annals of Mathematical Statistics* ***41*** *495–502.*

KOLTCHINSKII, V. *and* GINÉ, E. *(2000). Random matrix approximation of spectra of integral operators. Bernoulli 113–167.*

KOUL, H., SUSARLA, V. *and* VAN RYZIN, J. *(1981). Regression analysis with randomly right-censored data. The Annals of Statistics 1276–1288.*

LANCKRIET, G., CRISTIANINI, N., BARTLETT, P., GHAOUI, L. *and* JORDAN, M. *(2004a). Learning the kernel matrix with semidefinite programming. The Journal of Machine Learning Research* ***5*** *27–72.*

LANCKRIET, G., DENG, M., CRISTIANINI, N., JORDAN, M., NOBLE, W. ET AL. *(2004b). Kernel-based data fusion and its application to protein function prediction in yeast. In Proceedings of the Pacific Symposium on Biocomputing, vol. 9. World Scientific Singapore.*

LI, H. *and* LUAN, Y. *(2003). Kernel cox regression models for linking gene expression profiles to censored survival data. Pac Symp Biocomput 65–76.*

LI, Y., WILLER, C., SANNA, S. *and* ABECASIS, G. *(2009). Genotype imputation. Annu Rev Genomics Hum Genet* ***10*** *387–406.*

LI, Y., WILLER, C. J., DING, J., SCHEET, P. *and* ABECASIS, G. R. *(2010). MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol* ***34*** *816–834.*

LIU, D., GHOSH, D. *and* LIN, X. *(2008). Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models. BMC Bioinformatics* ***9*** *292.*

LIU, D., LIN, X. *and* GHOSH, D. *(2007). Semiparametric regression of multidimensional genetic pathway data: least-squares kernel machines and linear mixed models. Biometrics* ***63*** *1079–88.*

LIU, Z., CHEN, D., TAN, M., JIANG, F. *and* GARTENHAUS, R. B. *(2010). Kernel based methods for accelerated failure time model with ultra-high dimensional data. BMC Bioinformatics* ***11*** *606.*

LUAN, Y. *and* LI, H. *(2008). Group additive regression models for genomic data analysis. Biostatistics* ***9*** *100–113.*

LUCA, D., RINGQUIST, S., KLEI, L., LEE, A. B., GIEGER, C., WICHMANN, H.-E., SCHREIBER, S., KRAWCZAK, M., LU, Y., STYCHE, A., DEVLIN, B., ROEDER, K. *and* TRUCCO, M. *(2008). On the use of general control samples for genome-wide association studies: genetic matching highlights causal variants. Am J Hum Genet* ***82*** *453–63.*

MARCHINI, J. *and* HOWIE, B. *(2010). Genotype imputation for genome-wide association studies. Nat Rev Genet* ***11*** *499–511.*

MCCARTHY, M. I., ABECASIS, G. R., CARDON, L. R., GOLDSTEIN, D. B., LITTLE, J., IOANNIDIS, J. P. A. *and* HIRSCHHORN, J. N. *(2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet* ***9*** *356–369.*

MEIER, L., VAN DE GEER, S. *and* BÜHLMANN, P. *(2008). The group lasso for logistic regression. Journal of the Royal Statistical Society: Series B (Statistical Methodology)* ***70*** *53–71.*

MIKA, S., SCHÖLKOPF, B., SMOLA, A., MÜLLER, K., SCHOLZ, M. *and* RÄTSCH, G. *(1999). Kernel pca and de-noising in feature spaces. Advances in neural information processing systems* ***11*** *536–542.*

MOSKVINA, V., CRADDOCK, N., HOLMANS, P., OWEN, M. J. *and* O'DONOVAN, M. C. *(2006). Effects of differential genotyping error rate on the type I error probability of case-control studies. Hum Hered* ***61*** *55–64.*

NAROD, S. *and* FOULKES, W. *(2004). Brca1 and brca2: 1994 and beyond. Nature Reviews Cancer* ***4*** *665–676.*

NYHOLT, D. *(2004). A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. The American Journal of Human Genetics* ***74*** *765–769.*

PATTERSON, N., PRICE, A. L. *and* REICH, D. *(2006). Population structure and eigenanalysis. PLoS Genet* ***2*** *e190.*

POLLARD, D. *(1990). Empirical processes: theory and applications. In NSF-CBMS regional conference series in probability and statistics. JSTOR.*

PRICE, A. L., PATTERSON, N. J., PLENGE, R. M., WEINBLATT, M. E., SHADICK, N. A. *and* REICH, D. *(2006). Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet* ***38*** *904–909.*

PURCELL, S., NEALE, B., TODD-BROWN, K., THOMAS, L., FERREIRA, M. A. R., BENDER, D., MALLER, J., SKLAR, P., DE BAKKER, P. I. W., DALY, M. J. *and* SHAM, P. C. *(2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet* ***81*** *559–575.*

QI, L., CORNELIS, M. C., KRAFT, P., STANYA, K. J., LINDA KAO, W. H., PANKOW, J. S., DUPUIS, J., FLOREZ, J. C., FOX, C. S., PARE, G., SUN, Q., GIRMAN, C. J., LAURIE, C. C., MIREL, D. B., MANOLIO, T. A., CHASMAN, D. I., BOERWINKLE, E., RIDKER, P. M., HUNTER, D. J., MEIGS, J. B., LEE, C.-H., HU, F. B. *and* VAN DAM, R. M. *(2010). Genetic variants at 2q24 are associated with susceptibility to type 2 diabetes. Hum Mol Genet* ***19*** *2706–2715.*

R DEVELOPMENT CORE TEAM *(2009). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.*
URL `http://www.R-project.org`

RASMUSSEN, C. *and* WILLIAMS, C. *(2006). Gaussian processes for machine learning. 2006. The MIT Press, Cambridge, MA, USA* ***38*** *715–719.*

REIS-FILHO, J. *and* PUSZTAI, L. *(2011). Gene expression profiling in breast cancer: classification, prognostication, and prediction. The Lancet* ***378*** *1812–1823.*

RITOV, Y. *(1990). Estimation in a linear regression model with censored data. The Annals of Statistics 303–328.*

SCHOLKOPF, B. *and* SMOLA, A. *(2002). Learning with kernels. MIT Press Cambridge, Mass.*

SCHÖLKOPF, B., SMOLA, A. *and* MÜLLER, K. *(1998). Nonlinear component analysis as a kernel eigenvalue problem. Neural computation* ***10*** *1299–1319.*

SCOTT, L. J., MOHLKE, K. L., BONNYCASTLE, L. L., WILLER, C. J., LI, Y., DUREN, W. L., ERDOS, M. R., STRINGHAM, H. M., CHINES, P. S., JACKSON, A. U., PROKUNINA-OLSSON, L., DING, C.-J., SWIFT, A. J., NARISU, N., HU, T., PRUIM, R., XIAO, R., LI, X.-Y., CONNEELY, K. N., RIEBOW, N. L., SPRAU, A. G., TONG, M., WHITE, P. P., HETRICK, K. N., BARNHART, M. W., BARK, C. W., GOLDSTEIN, J. L., WATKINS, L., XIANG, F., SARAMIES, J., BUCHANAN, T. A., WATANABE, R. M., VALLE, T. T., KINNUNEN, L., ABECASIS, G. R., PUGH, E. W., DOHENY, K. F., BERGMAN, R. N., TUOMILEHTO, J., COLLINS, F. S. *and* BOEHNKE, M. *(2007). A genome-wide association study of type 2 diabetes in finns detects multiple susceptibility variants. Science* ***316*** *1341–1345.*

SEBASTIANI, P., SOLOVIEFF, N., PUCA, A., HARTLEY, S., MELISTA, E., ANDERSEN, S., DWORKIS, D., WILK, J., MYERS, R., STEINBERG, M., MONTANO, M., BALDWIN, C. *and* PERLS, T. *(2010). Genetic signatures of exceptional longevity in humans. Science .*

SONNENBURG, S., RÄTSCH, G., SCHÄFER, C. *and* SCHÖLKOPF, B. *(2006). Large scale multiple kernel learning. The Journal of Machine Learning Research* ***7*** *1531–1565.*

SOTIRIOU, C., WIRAPATI, P., LOI, S., HARRIS, A., FOX, S., SMEDS, J., NORDGREN, H., FARMER, P., PRAZ, V., HAIBE-KAINS, B. ET AL. *(2006). Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. Journal of the National Cancer Institute* ***98*** *262.*

TSIATIS, A. *(1990). Estimating regression parameters using linear rank tests for censored data. The Annals of Statistics 354–372.*

UNO, H., CAI, T., PENCINA, M. J., D'AGOSTINO, R. B. *and* WEI, L. J. *(2011). On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Stat Med* ***30*** *1105–17.*

VAN DER VAART, A. *(1994). Weak convergence of smoothed empirical processes. Scandinavian Journal of Statistics 501–504.*

WANG, H. *and* LENG, C. *(2007). Unified lasso estimation by least squares approximation. Journal of the American Statistical Association* ***102*** *1039–1048.*

WANG, S., NAN, B., ZHU, N. *and* ZHU, J. *(2009). Hierarchically penalized cox regression with grouped variables. Biometrika* ***96*** *307–322.*

WANG, Y., KLIJN, J., ZHANG, Y., SIEUWERTS, A., LOOK, M., YANG, F., TALANTOV, D., TIMMERMANS, M., MEIJER-VAN GELDER, M., YU, J. ET AL. *(2005). Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. The Lancet* ***365*** *671–679.*

WEI, L. J. *(1992). The accelerated failure time model: a useful alternative to the cox regression model in survival analysis. Stat Med* ***11*** *1871–9.*

WEI, Z. *and* LI, H. *(2007). Nonparametric pathway-based regression models for analysis of genomic data. Biostatistics* ***8*** *265–284.*

WELLCOME TRUST CASE CONTROL CONSORTIUM *(2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature* ***447*** *661–678.*

WRENSCH, M., JENKINS, R. B., CHANG, J. S., YEH, R.-F., XIAO, Y., DECKER, P. A., BALLMAN, K. V., BERGER, M., BUCKNER, J. C., CHANG, S., GIANNINI, C., HALDER, C., KOLLMEYER, T. M., KOSEL, M. L., LACHANCE, D. H., MCCOY, L., O'NEILL, B. P., PATOKA, J., PICO, A. R., PRADOS, M., QUESENBERRY, C., RICE, T., RYNEARSON, A. L., SMIRNOV, I., TIHAN, T., WIEMELS, J., YANG, P. *and* WIENCKE, J. K. *(2009). Variants in the CDKN2B and RTEL1 regions are associated with high-grade glioma susceptibility. Nat Genet* ***41*** *905–908.*

XU, X., TIAN, L. *and* WEI, L. *(2003). Combining dependent tests for linkage or association across multiple phenotypic traits. Biostatistics* ***4*** *223–229.*

YUAN, M. *and* LIN, Y. *(2006). Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology)* ***68*** *49–67.*

ZAYKIN, D. V., ZHIVOTOVSKY, L. A., WESTFALL, P. H. *and* WEIR, B. S. *(2002). Truncated product method for combining p-values. Genet Epidemiol* ***22*** *170–85.*

ZENG, D. *and* LIN, D. *(2007). Efficient estimation for the accelerated failure time model. Journal of the American Statistical Association* ***102*** *1387–1396.*

ZHANG, H. *and* LU, W. *(2007). Adaptive lasso for cox's proportional hazards model. Biometrika* ***94*** *691–703.*

ZHAO, S. *and* LI, Y. *(2011). Principled sure independence screening for cox models with ultra-high-dimensional covariates. Journal of Multivariate Analysis .*

ZHUANG, J. J., ZONDERVAN, K., NYBERG, F., HARBRON, C., JAWAID, A., CARDON, L. R., BARRATT, B. J. *and* MORRIS, A. P. *(2010). Optimizing the power of genome-wide association studies by using publicly available reference samples to expand the control group. Genet Epidemiol* ***34*** *319–326.*